Consistent Estimation in Household Surveys

Submitted in partial fulfillment of the requirements for the degree

Dr. rer. pol.

to the
Department IV
at Trier University

submitted by

Diplom Volkswirtin Anne Konrad, M.Sc.
Kloschinskystraße 12, 54292 Trier
born 06.08.1984 in Berlin

Supervisors:
Prof. Dr. Ralf Münnich (Trier University)
Ass. Prof. Dr. Yves Berger (University of Southampton)

March 2019

Contents

Acknowledgements

First of all, I am profoundly grateful to my first supervisor Ralf Münnich for his support. He created a comprehensive infrastructure at the Department of Economic and Social Statistics, which enables us to work in a very productive environment.

I am also very grateful to my second supervisor Yves Berger for all his valuable and constructive comments on my work and in general on scientific research. I really enjoyed this interchange.

I have to thank the RIFOSS project with the Federal Statistical Office of Germany, which provided the funding for my position as research associate at the University of Trier.

For the warmly, cooperative and productive working atmosphere and discussions I sincerely thank all of my colleagues at the Department of Economic and Social Statistics.

My special gratitude is due to my family who always trust in me and their continuously support in all life situations.

Finally, I owe my loving thanks to my boyfriend Matthias Braband. He has been my greatest source of perpetual support, and encouragement throughout all these years.

German Summary

Diese Dissertation beschäftigt sich mit konsistenten Schätzungen in Haushaltsstichproben. Haushaltsstichproben werden häufig als Klumpenstichprobe realisiert, das heißt auf der ersten Stufe werden die Haushalte gezogen und auf der zweiten Stufe die Personen innerhalb eines Haushaltes. Dabei stellt sich die Frage, inwiefern konsistente Schätzungen auf Personen- und Haushaltsebene erreicht werden können. Beispielsweise sollte das geschätzte Gesamtein- kommen aller Personen zu dem geschätzten Gesamteinkommen aller Haushalte gleich sein. Die Forderung nach konsistenten Schätzungen spielt eine wichtige Rolle in der amtlichen Statistik und ist als ein Prinzip im Verhaltenskodex für europäische Statistiken (European Code of Practice) (Eurostat, 2011, Prinzip 14) verankert. In der bisherigen Praxis verwenden die Statistischen Ämter die integrierte Gewichtung. Hierbei werden die individuellen Personenmerkmale durch den Haushaltsmittelwert ersetzt. Aufgrund der identischen Ausprägungen der Hilfsmerkmale sind auch die resultierenden Gewichte aller Personen innerhalb eines Haushaltes gleich. Auf Haushaltsebene findet keine separate Berechnung der Gewichte statt, stattdessen erhalten die Haushalte das Gewicht der Haushaltsmitglieder. Durch diese Gleichheit der Gewichte werden konsistente Schätzungen auf Personen- und Haushaltsebene garantiert. Jedoch gehen aufgrund der erzwungenen Gleichheit der Gewichte die individuellen Muster der Personen verloren, ebenso wird die Heterogenität der Haushalte nicht beachtet. Die Motivation dieser Dissertation ist es daher, die Auswirkungen dieser erzwungenen Gleichheit der Gewichte im integrierten Ansatz zu untersuchen sowie alternative Gewichtungsstrategien vorzuschlagen.

Kapitel 1 betont die Relevanz von konsistenten Schätzungen auf Personen- und Haushaltsebene. Wichtige Konzepte der Survey Statistik werden in Kapitel 2 eingeführt. Der Fokus liegt hierbei auf dem GREG Schätzer und Klumpenstichproben.

In Kapitel 3 verdeutlichen wir, dass als Konsequenz gleicher Gewichte im integrierten Ansatz die Anzahl möglicher Ausprägungen der Hilfsmerkmale ansteigt, die Within-Varianz der Haushalte ignoriert wird und Ecological Fallacy auftreten kann. Eine Simulationsstudie basierend auf einer realitätsnahen synthetischen Grundgesamtheit zeigt, dass diese Auswirkungen zu variableren Gewichten und weniger effizienten Punkt- und Varianzschätzungen in kleineren Stichprobenumfängen verglichen mit einem naiven GREG Schätzer führt.

Um diese Auswirkungen zu vermeiden, schlagen wir alternative Gewichtungsverfahren vor, welche sowohl konsistente Schätzungen gewährleisten als auch individuelle Gewichte innerhalb eines Haushaltes zulassen (Kapitel 4). Die Idee der alternativen Gewichtungsverfahren ist es die Konsistenzbedingungen auf Variablen zu beschränken, die sowohl im Personen- als auch im Haushaltsdatensatz vorkommen. Diese gemeinsamen Variablen werden als zusätzliche

Hilfsmerkmale in die Schätzung auf Personen- und Haushaltsebene aufgenommen. Damit wird Konsistenz direkter und nur für die relevanten Merkmale erreicht, anstatt indirekt durch den Zwang gleicher Gewichte für alle Personen eines Haushaltes. Entscheidende Vorteile unserer Gewichtungsstrategien sind, neben variablen Personengewichten, dass die originalen Hilfsmerkmale verwendet werden können und dass auf Personen- und Haushaltsebene separate Gewichtungsmodelle implementiert werden können, wodurch die Flexibilität in der Variablenselektion erhöht ist.

Um die Effekte der Konsistenzbedingung abschätzen zu können, vergleichen wir die asymptotischen Varianzen eines integrierten und eines naiven GREG Schätzers (Kapitel 5). Aus einem solchen Effizienzvergleich schlussfolgern Steel and Clark (2007), dass der integrierte GREG Schätzer einem naiven GREG Schätzer vorzuziehen ist. Da diese Schlussfolgerung unserer Argumentation in den vorangegangen Kapiteln und unseren Simulationsergebnissen widerspricht, zeigen wir zunächst einige Schwächen in dem Beweis von Steel and Clark (2007) auf. Unter anderem vernachlässigen sie den Interzept im integrierten Modell. Anschließend leiten wir einen eigenen Effizienzvergleich zwischen einem naiven und einem integrierten GREG Schätzer her. Eine Herausforderung besteht unter anderem darin, dass die zu vergleichenden Schätzer unterschiedlicher Dimensionen sind. Um dieses Problem zu lösen, zerlegen wir die asymptotische Varianz des integrierten GREG Schätzers in die Varianz eines reduzierten Schätzers ohne Interzept, der in der Dimensionen vergleichbar ist mit einem naiven GREG Schätzer und einen Anpassungsterm, der die Effekte des Interzepts, welche durch den reduzierten Schätzer vernachlässigt werden, erfasst. Anschließend schlagen wir für unsere Zerlegung ein weiteres Anwendungsfeld in der Variablenselektion im Bereich der Ökonometrie oder Survey Statistik vor.

In Kapitel 6 untersuchen wir die Varianzformel von GREG Schätzern in Klumpenstichproben. Dabei zeigen wir auf, dass für einen GREG Schätzer, der auf Personenebene modelliert ist, ein Trade-off zwischen der Optimalitätsbedingung und der Modellierungsebene besteht. Als Abhilfe schlagen wir einen Hybridschätzer vor, welcher zwischen der Optimalitätsbedingung und der Modellierung auf Personenebene abwägt.

Mit diesen Themen adressiert die Dissertation sowohl die praktische Anwendung in der amtlichen Statistik als auch theoretische Überlegungen zur Effizienzbetrachtung in der Survey Statistik. Jedes Kapitel schließt mit einer Simulationsstudie ab, um zuvor aufgestellte theoretische Überlegungen zu validieren.

List of Figures

List of Tables

List of Symbols

Symbols and Notation

U	Finite population of size N
U_p	Finite population of persons of size N
U_h	Finite population of households of size M
U_g	Finite population of persons within household g of size N_g
s	Sample of size n
s_p	Sample of persons of size n
s_h	Sample of households of size m
s_c	Combined sample of person and households with $s = s_p \cup s_h$
$\mathcal{S}$	Set of all possible samples
$p(*)$	Sampling design of *
π_*	First-order inclusion probability of $*$
π_{*+}	Second-order inclusion probability of $*$ and $+$
$\mathbf{\Pi}_*$	Diagonal matrix with inclusion probabilities of $*$
d_*	Design weight of *
$\boldsymbol{d}_*$	Vector of design weights of $*$
w_*	Weight of $*$
a	Positive constant
$\boldsymbol{Y}$	Random vector
$\boldsymbol{y}$	Vector of n realizations of the random vector $\boldsymbol{Y}$
θ	Unknown population parameter
$\hat{\theta}$	Estimator of the unknown population parameter
$\mathrm{E}(*)$	Expected value of $*$
$\mathrm{V}(*)$	Variance of $*$
$\mathrm{Cov}(*)$	Covariance matrix of $*$
$\mathrm{Cor}(*)$	Correlation matrix of $*$
$\mathrm{Bias}(*)$	Bias of $*$
$\mathrm{MSE}(*)$	Mean squared error of $*$
y_i, y_g	Value of the variable of interest of person i or household g
T_{y_p}	Unknown population total at the person level
T_{y_h}	Unknown population total at the household level
$\hat{T}^*_{y_p}$	Estimator of method $*$ for the unknown total T_{y_p} at the person level
$\hat{T}^*_{y_h}$	Estimator of method $*$ for the unknown total T_{y_h} at the household level
$\tilde{\boldsymbol{T}}_{\boldsymbol{c}}$	Estimator for the unknown common variable totals
R^2	Coefficient of determination

ξ	Linear regression model
$\boldsymbol{\beta}$	Population regression coefficient
$\hat{\boldsymbol{B}}$	Least squares estimate for the unknown population regression coefficient β
ϵ_*	Unobserved random error of $*$
v_*	Known variance parameter of $*$ with $v_* > 0$
R_*	Residual of $*$ at the population level
r_*	Estimated residual of $*$
$f(\cdot)$	Real-valued function
!	Factorial
$\doteq$	Limit value approach
α	Positive scale factor
$\boldsymbol{h}$	Constant vector
$G(\cdot)$	Distance function
$g(\cdot)$	Derivative of distance function $G(\cdot)$
$F(\cdot)$	Inverse function
$\boldsymbol{\lambda}$	Lagrange multipliers
L	Lower bound
U	Upper bound
M	Residual maker matrix
V_*	Variance component $*$
$\boldsymbol{Q}$	Weighting matrix of dimension $L \times L$

Indices

N	Number of persons in the population
n	Number of persons in the sample
M	Number of households in the population
m	Number of households in the sample
Q	Number of the person-level auxiliaries $\boldsymbol{x_i}$
L	Number of common variables
K	Number of the household-level auxiliaries $\boldsymbol{a_g}$
R	Number of Monte-Carlo replicates
i, j	Index of person $i, j \in s_p$
g, k	Index of the households $g, k \in s_h$
t	Index of the units $t \in s_c$
q	Index of auxiliary variable with $q = 1, \ldots, Q$
q'	Index of person-level auxiliary variables with $q' = 2, \ldots, Q$
k	Index of the household-level auxiliary variables with $k = 1, \ldots, K$
l	Index of the common variables with $l = 1, \ldots, L$
r	Index of Monte-Carlo replicates with $r = 1 \ldots, R$

Variables

$\boldsymbol{x}_i$	Vector of the auxiliary variables of person i of dimension Q
$\bar{\boldsymbol{x}}_i$	Vector of the mean values of the auxiliary variables of person i of dimension Q
$\bar{\boldsymbol{x}}_{i,-q}$	Vector of the mean values of the auxiliary variables of person i without the q-th element of dimension $(Q-1)$
$\boldsymbol{u}_i$	Instrumental variables of person i of dimension Q
$\boldsymbol{x}_g^\circ$	Integrated auxiliary vector of household g of dimension $Q+1$
$\bar{\boldsymbol{x}}_i^\circ$	Integrated auxiliary vector of person i of dimension $Q+1$
$\boldsymbol{a}_g$	Auxiliary variable vector of household g of dimension K
$\boldsymbol{x}_g$	Auxiliary variable vector aggregated per household g of dimension Q
$\boldsymbol{c}_i$	Common variable vector of person i of dimension L
$\boldsymbol{c}_g$	Common variable vector of household g of dimension L
$\boldsymbol{z}_i$	Additional explanatory variables of person i of dimension K (only in Section 5.3)
$\gamma_{t,p}$	Combined variable of interest of unit t at the person level
$\gamma_{t,h}$	Combined variable of interest of unit t at the household level
$\boldsymbol{\zeta}_t$	Vector of the combined auxiliary and common variables of unit t of dimension $Q+K+L$
$\boldsymbol{\delta}_t$	Vector of the combined auxiliary variables of unit t of dimension $Q+K$
$\boldsymbol{\kappa}_t$	Vector of the combined common variables of unit t of dimension L
$\boldsymbol{X}$	Matrix of the auxiliary variables at the person level of dimension $(n \times Q)$
$\boldsymbol{A}$	Matrix of the auxiliary variables at the household level of dimension $(m \times K)$
$\boldsymbol{C}_p$	Matrix of the common variables at the person level of dimension $(n \times L)$
$\boldsymbol{C}_h$	Matrix of the common variables at the household level of dimension $(m \times L)$
$\boldsymbol{Z}$	Matrix of the combined auxiliary and common variables of dimension $(n+m) \times (Q+K+L)$

Coefficients and Residuals in the Alternative Weighting Approaches (Chapter 4)

$\hat{\boldsymbol{D}}_{\boldsymbol{x}}$, $\hat{\boldsymbol{D}}_{\boldsymbol{c}}$ Coefficient vectors of the second proposed person-level estimator $\hat{T}^{\text{WA2}}_{y_p} = \hat{T}^{HT}_{y_p} + \hat{\boldsymbol{D}}_{\boldsymbol{x}}^{T}(\boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}^{\boldsymbol{HT}}_{\boldsymbol{x}}) + \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c}^*_{\boldsymbol{p}}} - \hat{\boldsymbol{T}}^{\boldsymbol{HT}}_{\boldsymbol{c}})$

$\hat{\boldsymbol{B}}_{\boldsymbol{x}}$ Coefficient vector of the naïve person-level GREG estimator $\hat{T}^{\text{GREG}}_{y_p} = \hat{T}^{\text{HT}}_{y_p} + \hat{\boldsymbol{B}}_{\boldsymbol{x}}^{T}(\boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{x}})$

$\hat{\boldsymbol{F}}_{\boldsymbol{x}}$ Coefficient vector of the estimator of the common variable totals at the person level $\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} = \hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{c_p}} + \hat{\boldsymbol{F}}_{\boldsymbol{x}}^{T}(\boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{x}})$

$\hat{\boldsymbol{E}}_{\boldsymbol{a}}$, $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ Coefficient vectors of the first proposed household-level estimator $\hat{T}^{\text{WA1}}_{y_h} = \hat{T}^{\text{HT}}_{y_h} + \hat{\boldsymbol{E}}_{\boldsymbol{a}}^{T}(\boldsymbol{T}_{\boldsymbol{a}} - \hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{a}}) + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{c_p}})$

$\hat{\boldsymbol{B}}_{\boldsymbol{a}}$ Coefficient vector of the naïve household-level GREG estimator $\hat{T}^{\text{GREG}}_{y_h} = \hat{T}^{\text{HT}}_{y_h} + \hat{\boldsymbol{B}}_{\boldsymbol{a}}^{T}(\boldsymbol{T}_{\boldsymbol{a}} - \hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{a}})$

$\hat{\boldsymbol{F}}_{\boldsymbol{a}}$ Coefficient vector of the estimator of the common variable totals at the household level $\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}} = \hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{c_h}} + \hat{\boldsymbol{F}}_{\boldsymbol{a}}^{T}(\boldsymbol{T}_{\boldsymbol{a}} - \hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{a}})$

$\hat{\boldsymbol{\Psi}}_{\boldsymbol{p}}$ Person-level coefficient vector of the combined GREG estimator $\hat{T}^{\text{ZIE}}_{\gamma_p} = \hat{T}^{\text{HT}}_{\gamma_p} + \hat{\boldsymbol{\Psi}}_{\boldsymbol{p}}^{T}(\boldsymbol{T}_{\zeta} - \hat{\boldsymbol{T}}^{\text{HT}}_{\zeta})$

$\hat{\boldsymbol{\Psi}}_{\boldsymbol{h}}$ Household-level coefficient vector of the combined GREG estimator $\hat{T}^{\text{ZIE}}_{\gamma_h} = \hat{T}^{\text{HT}}_{y_h} + \hat{\boldsymbol{\Psi}}_{\boldsymbol{h}}^{T}(\boldsymbol{T}_{\zeta} - \hat{\boldsymbol{T}}^{\text{HT}}_{\zeta})$

$\hat{\boldsymbol{D}}_{\delta}$, $\hat{\boldsymbol{D}}_{\kappa}$ Coefficient vectors of the combined GREG estimator at the person level $\hat{T}^{\text{ZIE}}_{\gamma_p} = \hat{T}^{\text{HT}}_{\gamma_p} + \hat{\boldsymbol{D}}_{\delta}^{T}(\boldsymbol{T}_{\delta} - \hat{\boldsymbol{T}}^{\text{HT}}_{\delta}) + \hat{\boldsymbol{D}}_{\kappa}^{T}(\boldsymbol{0} - \hat{\boldsymbol{T}}^{\textbf{HT}}_{\boldsymbol{\kappa}})$

$\hat{\boldsymbol{B}}_{\delta}$ Coefficient vector of the naïve combined model $\gamma^p_t = \hat{\boldsymbol{B}}_{\delta}^{T}\boldsymbol{\delta}_{\boldsymbol{t}} + \boldsymbol{r}^{\boldsymbol{B}_{\delta}}_{\boldsymbol{t}}$

$\hat{\boldsymbol{F}}_{\delta}$ Coefficient vector of the auxiliary combined model $\boldsymbol{\kappa}_t = \hat{\boldsymbol{F}}_{\delta}^{T}\boldsymbol{\delta}_{\boldsymbol{t}} + \boldsymbol{r}^{\boldsymbol{F}_{\delta}}_{\boldsymbol{t}}$

$\hat{\boldsymbol{E}}_{\delta}$, $\hat{\boldsymbol{E}}_{\kappa}$ Coefficient vector of the combined GREG estimator at the household level $\hat{T}^{\text{ZIE}}_{\gamma_h} = \hat{T}^{\text{HT}}_{\gamma_h} + \hat{\boldsymbol{E}}_{\delta}^{T}(\boldsymbol{T}_{\delta} - \hat{\boldsymbol{T}}^{\text{HT}}_{\delta}) + \hat{\boldsymbol{E}}_{\kappa}^{T}(\boldsymbol{0} - \hat{\boldsymbol{T}}^{\textbf{HT}}_{\boldsymbol{\kappa}})$

$r_i^{B_x}$ Residual of the naïve model at the person level with $r_i^{B_a} = y_i - \hat{\boldsymbol{B}}_{\boldsymbol{x}}^{T}\boldsymbol{x}_{\boldsymbol{i}}$

$\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F}_x}$ Residual of the auxiliary model at the person level with $\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F}_x} = \boldsymbol{c}_{\boldsymbol{i}} - \hat{\boldsymbol{F}}_{\boldsymbol{x}}^{T}\boldsymbol{x}_{\boldsymbol{i}}$

$r_g^{B_a}$ Residual of the naïve model at the household level with $r_g^{B_a} = y_g - \hat{\boldsymbol{B}}_{\boldsymbol{a}}^{T}\boldsymbol{a}_{\boldsymbol{g}}$

$\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F}_a}$ Residual of the auxiliary model at the household level with $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F}_a} = \boldsymbol{c}_{\boldsymbol{g}} - \hat{\boldsymbol{F}}_{\boldsymbol{a}}^{T}\boldsymbol{a}_{\boldsymbol{g}}$

Models and Residuals in the Efficiency Comparison (Chapter 5)

$y_i = H_x x_i + H_{\bar{x}} \bar{x}_i + r_i^H$	Model for explanation of the FWL theorem
$y_i = B_p x_i + r_i^{B_p}$	Model for explanation of the FWL theorem
$\bar{x}_i = B_x x_i + r_i^{B_x}$	Model for explanation of the FWL theorem
$y_i = \boldsymbol{B_p}^T \boldsymbol{x_i} + r_i^{B_p}$	Person-level model
$y_i = B_p x_{i1} + r_i^{B_p}$	Simple person-level model
$y_g = \boldsymbol{B_h^{\circ}}^T \boldsymbol{x_g^{\circ}} + r_g^{B_h^{\circ}}$	Integrated household-level model
$= B_{x_0}^{\circ} x_{g0} + \boldsymbol{B_x^{\circ}}^T \boldsymbol{x_g} + r_g^{B^{\circ}}$	
$y_g = \boldsymbol{B_h^{\circ}}^T \boldsymbol{x_g^{\circ}} + r_g^{B_h^{\circ}}$	Simple integrated household-level model
$= B_{x_0}^{\circ} x_{g0} + B_{x_1}^{\circ} x_{g1} + r_g^{B^{\circ}}$	
$y_g = \boldsymbol{B_h}^T \boldsymbol{x_g} + r_g^{B_h}$	Reduced household-level model
$y_g = B_h x_{g1} + r_g^{B_h}$	Simple reduced household-level model
$y_i = \boldsymbol{B_h}^T \boldsymbol{\bar{x}_i} + r_i^{B_h}$	Reduced person-level model
$x_{g0} = \boldsymbol{F_x}^T \boldsymbol{x_g} + r_g^{F_x}$	Auxiliary model for the decomposition
$x_{g0} = F_{x_1} x_{g1} + r_g^{F_{x_1}}$	Simple auxiliary model for the decomposition
$y_i = D_{x_1} x_{i1} + \boldsymbol{D_x}^T \boldsymbol{x_i} + \boldsymbol{D_{\bar{x}}}^T \boldsymbol{\bar{x}_i} + r_i^D$	Overlap model
$y_i = D_{x_1} x_{i1} + D_{x_2} x_{i2} + D_{\bar{x}_2} \bar{x}_{i2} + r_i^D$	Simple overlap model
$\bar{x}_i = \boldsymbol{B_x}^T \boldsymbol{x_i} + r_i^{B_{\bar{x}}}$	Auxiliary model for deriving the relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$
$x_i = \boldsymbol{B_{\bar{x}}}^T \boldsymbol{\bar{x}_i} + r_i^{B_x}$	Auxiliary model for deriving the relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$
$x_{iq'} = \boldsymbol{B_{\bar{x}'_q}}^T \boldsymbol{\bar{x}_i} + r_i^{B_{\bar{x}'_q}}$	Model to prove the form of $\boldsymbol{B_{\bar{x}}}$
$x_{iq'} = \boldsymbol{H_{\bar{x}}}^T \boldsymbol{\bar{x}_{i,-q}} + r_i^{H_{\bar{x}}}$	Model to prove the form of $\boldsymbol{B_{\bar{x}}}$
$\bar{x}_{iq} = \boldsymbol{K_{\bar{x}}}^T \boldsymbol{\bar{x}_{i,-q}} + r_i^{K_{\bar{x}}}$	Model to prove the form of $\boldsymbol{B_{\bar{x}}}$
$x_{iq'} = \boldsymbol{\tilde{H}_{\bar{x}}}^T \boldsymbol{\bar{x}_{i,-q}} + r_i^{\tilde{H}_{\bar{x}}}$	Model to prove the form of $\boldsymbol{B_{\bar{x}}}$
$\bar{x}_{i1} = \boldsymbol{\tilde{K}_{\bar{x}}}^T \boldsymbol{\bar{x}_{i,-q}} + r_i^{\tilde{K}_{\bar{x}}}$	Model to prove the form of $\boldsymbol{B_{\bar{x}}}$
$r_i^{B_p} = B_c \bar{x}_i + r_i^{B_c}$	Residual in the Steel and Clark (2007) approach
$\tilde{r}_g^{B_{x_0}^{\circ}} = y_g - B_{x_0}^{\circ} x_{g0}$	Pseudo-residual
$\tilde{r}_g^{B_{x_0}^{\circ} \cdot F_x} = y_g - B_{x_0}^{\circ} \cdot \boldsymbol{F_x}^T \boldsymbol{x_g}$	Pseudo-residual

List of Abbreviations

Abbreviations	
BLU estimator	Best linear unbiased estimator
deff	Design effect
EL	Empirical likelihood
FWL theorem	Frisch-Waugh-Lovell theorem
GLS	Generalized least squares
GREG estimator	Generalized regression estimator
HYB	Hybrid estimator
ID	Identifier
i.e.	id est
LFS	Labour Force Survey
MC	Monte-Carlo
MSE	Mean squared error
OLS	Ordinary least squares
PSU	Primary sampling unit
R	Programming software R
RB	Relative bias
RIS	Regional Income Survey
RRMSE	Relative root mean squared error
rsRB	Replicate-specific relative bias
SRS	Simple random sampling
SSR	Sum of squares regression
SSCS	Simple single-stage cluster sampling
SST	Sum of squares total
SSU	Secondary sampling unit

Estimators

approx	Approximated estimator
cal	Calibration estimator
calF	Calibration estimator following the functional approach
ES	Estevao and Särndal
GREG	Generalized regression estimator
HT	Horvitz-Thompson estimator
HYB	Hybrid estimator
INT	Integrated Weighting
LD	Lemaître and Dufour
MER	Merkouris
N	Nieuwenbroek
OPT	Optimal estimator
PERS	Person-level GREG estimator
RN	Renssen and Nieuwenbroek
SC	Steel and Clark
WA	Weighting Approach
ZIE	Zieschang

1 Relevance of Consistent Estimation in Household Surveys

Household surveys are an important source of socioeconomic and demographic data for scientific researchers from different fields as well as for political decision makers. Almost every country conducts household surveys. Such surveys are often drawn via cluster sampling, with households sampled at the first stage and persons selected at the second stage. The collected data provide information for estimation at both the person and the household level. However, consistent estimates are desirable in the sense that the estimated household-level totals should coincide with the estimated totals obtained at the person level.

In the literature, the terms coherence and consistency are often used synonymously. The term coherence tends to be more prevalent in the official statistics context, whereas consistency prevails in scientific research. Following Wallgren and Wallgren (2007, p. 219), "coherence refers to the fact that estimates from different surveys can be used together." On the other hand, consistency is defined, according to the glossary of statistical terms, as logical and numerical coherence (cf. OECD, 2007, p. 136). Särndal (2007) used the term consistency in the sense of being consistent with known totals. Therefore, the term consistency is used throughout this thesis because we are interested in numerical equal estimates of totals that are common to the person- and household-level data set and coherence does not necessarily imply full numerical consistency (cf. OECD, 2007, p. 120). It is important to note that here consistency does not refer to the property of convergence in probability. We return to this distinction in Section 2.2.2.

Integrated weighting introduced by Lemaître and Dufour (1987) is the current practice in official statistics to ensure consistent estimates. It produces one single weight for all persons within the same household by substituting the original person-level auxiliary variables by the corresponding household mean value. This single person-level weight is assigned one-to-one to the corresponding household. Thus, consistent estimates at the person and household level are ensured by the equality of the weights of the persons within a household and the household itself.

However, assigning equal weights to all household members completely ignores the individual differences between persons. Intuitively, for very volatile variables, such as income, the resulting estimates are significantly influenced when the same weights are used for all household members regardless of whether they are top earners, children, or inactive persons. Therefore, this thesis proposes alternative weighting approaches that ensure consistent estimates while overcoming the strict requirement of equal weights for all persons within the same household and the household itself. The underlying idea of our alternative weighting approaches is to constrain the consistency requirements to variables that are common to the person- and household-level data set. Thereby, consistency is ensured directly and solely for the relevant variables instead of indirectly by aggregating the individual information per household. With these alternative weighting approaches, we contradict the assumption prevailing in the literature that equal weights of persons within the same household and the household itself are necessary to ensure consistent estimates in household surveys (cf. Nieuwenbroek, 1993; Steel and Clark, 2007; Estevao and Särndal, 2006; Lavallée, 1995; Verma et al., 2006). Furthermore, to quantify the effect induced by the consistency requirements, this thesis provides an efficiency comparison of the asymptotic variances of a naïve and an integrated GREG estimator.

The contribution of this thesis to the literature is two-fold. First, the proposed weighting approaches serve as alternatives to integrated weighting to ensure consistent estimates in household surveys. Consistent estimation plays a vital role for official statistics, because survey users strive to obtain the same estimates for the same variable over different surveys (cf. Estevao and Särndal, 2006, p. 128). Consistency is also anchored as a principle in the European Statistics Code of Practice (cf. Eurostat, 2011, Principle 14), which sets the definition of quality criteria. Therefore, this thesis addresses a topic of high practical importance for official statistics. Second, we contribute to theoretical considerations in survey statistics. We propose a decomposition of asymptotic variances that allows us to compare the efficiency of models exhibiting different dimensions. Moreover, we initiate a discussion of the suitability of the variance formula for person-level GREG estimators under cluster sampling, because the initial level of modeling is ignored.

This thesis is organized into seven chapters.

Chapter 2 presents some basic preliminaries on survey statistics relevant for this thesis. The focus herein lies on cluster sampling and on GREG estimators.

Chapter 3 discusses integrated weighting as current practice in official statistics to ensure consistent estimates in household surveys. We explore the consequences of the strict requirement of equal weights and validate its effects on point and precision estimation by means of a simulation study. These theoretical and empirical results build the justification for our proposed alternative weighting approaches.

In Chapter 4, we introduce two alternative weighting approaches to guarantee consistent estimates. In contrast to integrated weighting, those approaches use the original auxiliary information and herein allow for different weights for the persons within a certain household. For this purpose, we adopt the idea of incorporating the common variables as additional auxiliaries known from the literature on multiple independent surveys. The two proposed alternative weighting approaches differ with respect to the implementation effort and the quality of the estimated common variables totals. The superiority of our proposed alternative weighting approaches compared to integrated weighting is validated by a subsequent simulation study.

Chapter 5, as the core part of the thesis, provides a theoretical comparison of the efficiency between the asymptotic variances of a person-level GREG estimator and an integrated GREG estimator. An efficiency comparison enables us to predict the factors on which the difference between the two variances depends. Steel and Clark (2007) claimed that under cluster sampling and for large samples, integrated weighting is more efficient than a person-level GREG estimator. Because this statement contradicts our argumentation and the results based on the simulation study given in Chapter 3, we detect some essential weaknesses on the given efficiency comparison. One main weakness is that Steel and Clark (2007) neglected the intercept in the integrated household-level model. Afterwards, we offer a correct efficiency comparison between a person-level GREG estimator and an integrated GREG estimator, circumventing the aforementioned insufficiencies. For this purpose, we contribute a procedure that decomposes the asymptotic variance of a GREG estimator into individual variance terms in order to separate the effect of a single variable. Finally, by extending the decomposition to multiple variables, we suggest further application fields.

In Chapter 6, we investigate the asymptotic variance of GREG estimators under cluster sampling. The corresponding formula implies a trade-off between person-level modeling and optimality. As a remedy, we introduce the hybrid GREG estimator as a compromise between optimality and person-level modeling.

Chapter 7 summarizes the findings of this thesis and draws a overall conclusion. In addition, we give an outlook for future research.

To facilitate the reading of this thesis, we introduce some general indications. Important concepts and terms appear in bold letters. Variables are written in `typewriter font`. Estimators and scenarios are characterized by capital letters. The main theoretical findings are presented as lemmas or results. To reinforce the comprehension, we present the calculation steps in a stepwise and very detailed manner. The correctness of individual calculation steps within the proofs in Chapter 5 are verified by the corresponding R code. We present our formulas in the more elaborate sum notation instead of in matrix notation, because the sum notation clearly indicates the level of estimation. Notation is consistent across chapters. Equations are numbered only if they are referenced later in the thesis.

2 Preliminaries in Survey Statistics

In scientific research, the need for statistical information is continuously increasing. Surveys gather statistical information about items such as total population sizes, unemployment rates, number of immigrants, poverty rates, or retail sales volumes. The German Federal Statistical Office is mandated by law by the Bundesstatistikgesetz (cf. Destatis, 2016) to provide statistical information. The advantage of drawing a random sample instead of a complete census of the whole population is that a sample (1) can provide reliable information at lower cost, (2) is less time-consuming, and (3) produces estimates that are often more accurate than those based on a census because the reduced amount of data to be collected enables greater care in the collecting process (cf. Lohr, 2009, p. 18).

Survey statistics supply methods to select random samples from a population (selection process) and use the sample data to compute estimates of unknown population parameters (estimation process). This chapter only briefly sketches preliminaries of the selection and estimation process to embed this thesis in the framework of survey statistics. In this context, only topics relevant for the thesis are addressed. The interested reader is referred to the quoted literature. The selection process and a general definition of probability sampling are outlined in Section 2.1. Section 2.2 presents the estimation process and introduces three distinct concepts of statistical inference as well as quality measures to evaluate different estimators. Because the thesis deals with generalized regression (GREG) estimators and cluster sampling, both topics are outlined as representatives of the estimation and selection process in more detail in Sections 2.3 and 2.4, respectively.

2.1 Selection Process

In the selection process, rules and operations define which units from the finite population are selected into the sample (cf. Kish, 1965, p. 4). In survey statistics, the randomness results from the sampling process. Therefore, it is fixed to which unit the observed values belong. In contrast, in econometrics, the realization of the values is stochastic. We start with introducing some basic notation and terminology. Consider a finite population $U = \{1, \ldots, i, \ldots, N\}$ of size N. It is assumed that N is known. If the population size is unknown in practice, it has to be estimated. A sample $s \subset U$ of size n with $s = \{1, \ldots, i, \ldots, n\}$ is selected from the population. A sample is called a probability sample if and only if the conditions of the following definition are fulfilled. The definition originates from Särndal et al. (1992, p. 8).

Definition 1. ***Probability Sampling***
Probability sampling is an approach to sample selection that satisfies certain conditions, which for the case of selecting elements directly from the population are described as follows:

1. *We can define the set of samples, $\mathcal{S} = \{s_1, \ldots, s_v\}$, that are possible to obtain with the sampling procedure.*

2. *A known probability of selection $p(s)$ is associated with each possible sample s.*

3. *The procedure gives every element in the population a nonzero probability of selection.*

4. *We select one sample by a random mechanism under which each possible s receives exactly the probability $p(s)$.*

For any sample satisfying these conditions, we can calculate the distribution of an estimator if repeatedly applied to the same population (cf. Cochran, 1977, p. 9). The probability $p(\cdot)$ referred in point 2 of Definition 1 is called sampling design. It determines the probability distribution on the set of all 2^N different samples of the finite population U (cf. Lehtonen and Pahkinen, 2004, p. 13). The probability referred to in point 3 is called first-order inclusion probability of unit i denoted as $\pi_i = Pr(i \in s) = \sum_{s \in U: i \in s} p(\cdot)$ (cf. Breidt and Opsomer, 2017, p. 190). In the design-based and model-assisted approach (defined in the following), statistical inference is based on the inclusion probabilities reflecting the design. The inverse of the inclusion probability is called design weight $d_i = \pi_i^{-1}$. The second-order inclusion probability is denoted by $\pi_{ij} = Pr(i, j \in s) = \sum_{s \in U: i, j \in s} p(\cdot)$. It gives the probability that both units i and j will be sampled. Note that $\pi_{ii} = \pi_i$. A sampling design is called measurable if $\pi_i > 0$ for all $i \in U$ and $\pi_{ij} > 0$ for all $i \neq j \in U$ (cf. Fuller, 2009, p. 11). This implies that all units in the population have a positive chance to be selected (cf. Hansen et al., 1953, p. 15).

The random mechanism referred to in point 4 of Definition 1 determines which units of the population listed in a **frame** are sampled. A sampling frame is defined as a list identifying all units in the population (cf. Lohr, 2009, p. 3). In practice, the compilation of such a list might be problematic. Kish (1965, pp. 53-59) gave an extensive overview of frame errors and remedies to reduce these errors.

On the basis of the concept of a frame, one can differentiate between direct element sampling and multistage sampling. In **direct element sampling**, a frame is available, and the population elements are the sampling elements (cf. Särndal et al., 1992, p. 61). Examples for direct element sampling designs are simple random sampling (SRS), systematic sampling, probability proportional-to-size and stratified sampling, Bernoulli sampling, and Poisson sampling. The interested reader is referred to Kish (1965), Särndal et al. (1992), Lohr (2009), and Cochran (1977) for a more detailed description of direct element sampling. In contrast, in **multistage sampling**, the population elements cannot be used as sampling elements, either because no sampling frame exists or because the population elements are widely scattered (cf. Särndal et al., 1992, p. 124). Section 2.4 discusses cluster sampling as one example of multistage sampling in more detail, because this thesis primarily deals with household surveys, which are frequently sampled by means of cluster sampling.

2.2 Estimation Process

In the estimation process, methods are applied to estimate finite population parameters such as means, totals, and ratios given the information of the selected units in the sample. In this section, three distinct concepts of statistical inference are briefly outlined (Section 2.2.1). Subsequently, we introduce certain quality measures to evaluate different estimators (Section 2.2.2). Finally, the well-known Horvitz-Thompson estimator is presented (Section 2.2.3).

2.2.1 Concepts of Statistical Inference

Statistical inference can be based on the design-based, design-assisted, and model-based approaches. The **design-based approach** was originated in the early work of Neyman (1934). In the design-based approach, statistical inference depends on the distribution generated by the sampling design while the population parameters are treated as fixed (cf. Lehtonen and Veijanen, 2009, p. 219). This implies that the randomness is induced by the probability to be sampled. No model is assumed for the underlying selection process. The design-based approach is followed by traditional sampling theory books, for example Hansen et al. (1953), Kish (1965), and Cochran (1977). A well-known design-based estimator is the Horvitz-Thompson estimator (introduced in Section 2.2.3). Sometimes the design-based approach is referred to as the **randomization-based approach**, for example in Kott (2005) or Lehtonen and Veijanen (2009).

To improve the efficiency of the estimators, the **model-assisted approach** postulates a statistical model of the unknown population parameter (cf. Breidt and Opsomer, 2017). It is closely related to the design-based approach and sometimes treated as a special case of it, for example in Särndal et al. (1978), Chambers (2011), and Little (2004). Design-based and model-assisted estimators expand the observed outcome values by survey weights. Models are used to assist the estimation process, whereby the resulting estimators are robust against model misspecification. Both design-based and model-assisted estimators are evaluated with design-based properties (introduced in Section 2.2.2) under repeated sampling from the fixed population with a given design. These properties do not depend on the correctness of the model (cf. Särndal et al., 1992, p. 239). A well-known model-assisted estimator is the GREG estimator presented in Section 2.3. The model-assisted approach is extensively discussed in Särndal et al. (1992).

The **model-based approach** is premised on the early work of Brewer (1963) and Royall (1970, 1976). Model-based estimators are motivated by the probability distribution of an assumed underlying statistical model. The outcome values of the non-sampled units are predicted on the basis of the model. The unknown population parameters are then estimated using both the observed and predicted outcome values of the sampled and the non-sampled units, respectively (cf. Valliant et al., 2000). In the model-based approach, randomness is induced because of the stochastic population structure of the realized population values as one outcome of a random variable (cf. Särndal et al., 1978). Statistical inference relies on the probability distribution of the assumed statistical model. Therefore, in contrast to the model-assisted approach, model-based statistical inference depends on the correctness of the assumed model. Model-based estimators are evaluated by model-based properties. With respect to the specification of the model, one can differentiate between superpopulation modeling and Bayesian modeling (cf. Little, 2004). Representatives of superpopulation modeling are Royall (1970) and Valliant et al. (2000). An overview of underlying models is given in Cassel et al. (1977). Representatives of the Bayesian approach include Ericson (1969, 1988), Binder (1982), Gosh and Meeden (1997), and Basu (2010). An important field of application for model-based inference is small-area estimation, which is useful if sample sizes are too small to produce reliable estimates and if additional information is available (see Rao, 2003 and Münnich et al., 2013, 2012a for more details). Because this thesis follows the design-based and model-assisted approaches, the model-based approach is not pursued hereinafter.

2.2.2 Quality Measures for Estimators

In the design-based and model-assisted approach, the sampled units are used to produce estimates for the unknown finite population parameters. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be a vector of n realizations out of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ of dimension N. A statistic $\hat{\theta} = f(\boldsymbol{y})$ using the realizations $\boldsymbol{y}$ to estimate the unknown population value θ is called an **estimator**. Before introducing certain estimators, we discuss how to evaluate the quality of different estimators at hand. A natural choice for a quality measure is how *close* an estimate is to the parameter to be estimated. It is desirable that in the long run the mean of the realizations of the estimator $\hat{\theta}$ equals the true population parameter θ. However, the realization of an estimator is a random value, whereas the true population parameter is fixed. As a remedy, closeness can be assessed in an expected or probabilistic sense (cf. Mittelhammer, 2013, p. 375). A measure describing the closeness in an expected sense is the **unbiasedness**. The bias of an estimator $\hat{\theta}$ of the unknown population parameter θ is defined as

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta,$$

where $E(\cdot)$ denotes the expected value (cf. Särndal et al., 1992, p. 40). The estimator is considered as unbiased if $E(\hat{\theta}) - \theta = 0$.

However, unbiased estimators can differ with respect to their sampling distributions around the true population parameter. This property is captured by the **efficiency** of an estimator. An unbiased estimator $\hat{\theta}$ of the unknown population parameter θ is called to be efficient if it has minimum variance in the class of unbiased estimators.

A quality measure combining the bias and efficiency is the mean squared error (MSE). The MSE of an estimator $\hat{\theta}$ of the unknown population parameter θ is defined as

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

The MSE can be decomposed into $\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$, where $V(\cdot)$ denotes the variance. This expression reflects the trade-off between variance and bias (cf. Schaich and Münnich, 2001, p. 190). For unbiased estimators, it follows that $\text{MSE}(\hat{\theta}) = \text{V}(\hat{\theta})$.

Unbiasedness, efficiency, and the MSE refer to finite sample properties, because the sample size of s is assumed to be fixed. In contrast, large sample properties relate to asymptotic theory with $n \rightarrow \infty$. The minimum requirement for an estimator is given by consistency. An estimator $\hat{\theta}_n$ of the unknown population parameter θ is called design-consistent, if

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

for all $\epsilon > 0$ holds (cf. Schaich and Münnich, 2001, p. 197). In other words, an estimator is design-consistent if its bias and variance tend to zero when the sample size increases (cf. Lehtonen and Veijanen, 2009, p. 222). For further discussion of large sample properties, such as sufficiency, asymptotic MSE, and asymptotic efficiency, the reader is referred to Fuller (2009, pp. 41) and Mittelhammer (2013, Section 7.3.3).

2.2.3 Horvitz-Thompson Estimator

A well-known design-based estimator incorporating the sampling design into the estimation process is the Horvitz-Thompson estimator (cf. Narain, 1951; Horvitz and Thompson, 1952). Let y_i be a non-random value of the variable of interest of unit i. For simplicity, we assume full response, that implies y_i is recorded for all $i \in s$. The objective is to estimate an unknown population total $T = \sum_{i \in U} y_i$. Further, more complex statistics can often be expressed as explicit functions of finite population totals, such as means, ratios, or regression coefficients. For the sake of convenience, the notation $\hat{T}_y$ refers to the estimator itself as well as to one realization of an estimate.

Result 1. ***The Horvitz-Thompson Estimator***
A design-unbiased estimator for the population total $T_y = \sum_{i \in U} y_i$ is given by the Horvitz-Thompson estimator

$$\hat{T}_y^{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

with variance

$$V(\hat{T}_y^{HT}) = \sum_{i \in U} \sum_{j \in U} \triangle_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j},$$

where $\triangle_{ij} = \pi_{ij} - \pi_i \pi_j$. Given that $\pi_{ij} > 0$ for all $i, j \in U$, an unbiased estimator of $V(\hat{T}_y^{HT})$ is given by

$$\hat{V}(\hat{T}_y^{HT}) = \sum_{i \in s} \sum_{j \in s} \frac{\triangle_{ij}}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{2.1}$$

Proof. See Särndal et al. (1992, p. 44). □

Breidt and Opsomer (2017) and Isaki and Fuller (1982) showed that the Horvitz-Thompson estimator is design-consistent under mild conditions. The design-unbiasedness is proven by Fuller (2002). The Horvitz-Thompson estimator is frequently used in official statistics, as it is simple to implement and robust for large sample fractions (cf. Münnich et al., 2012a, p. 39).

For equal probability sampling designs, the double sum in equation (2.1) vanishes. In the case of SRS with $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/N(N-1)$ the variance formula (2.1) simplifies to

$$\hat{V}_{SRS}(\hat{T}_y^{\mathrm{HT}}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y})^2 \tag{2.2}$$

with $\bar{y} = n^{-1}\sum_{i \in s} y_i$ as sample mean.

2.3 GREG Estimator

The efficiency of design-based estimators can be improved by incorporating auxiliary information in the estimation process. A widely used model-assisted estimator incorporating auxiliary information is the GREG estimator established by Hansen et al. (1953), Cassel et al. (1977), Särndal (1980), Isaki and Fuller (1982) as well as Wright (1983). Suppose that $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{iq}, \ldots, x_{iQ})^T$ is a vector containing Q auxiliaries. The corresponding totals $\boldsymbol{T_x} = (T_{x_1}, \ldots, T_{x_q}, \ldots, T_{x_Q})^T$ are known from censuses, registers, or other reliable sources. The GREG estimator relies on a linear regression model ξ that specifies the relationship between a variable of interest and the auxiliaries given by

$$y_i = \boldsymbol{x_i}^T\boldsymbol{\beta} + \epsilon_i \quad \text{for all} \quad i \in U \tag{2.3}$$

with $\boldsymbol{\beta}$ as population regression coefficient and ϵ_i as unobserved random error. Note that $E_\xi(\epsilon_i) = 0$, $V_\xi(\epsilon_i) = v_i\sigma^2$ and $\mathrm{E}_\xi(\epsilon_i\epsilon_j) = 0$ for all $i \neq j$. E_ξ, and V_ξ denote the expectation and the variance with respect to the model ξ. The variance parameter v_i with $v_i > 0$ has to be known and describes the residual pattern. Information about the variance component may be available from previous surveys. The choice $v_i = 1$ corresponds to the assumption of homoscedasticity. Homoscedasticity is often assumed in household surveys, because the variables of interest are particularly categorical (cf. Steel and Clark, 2007, p. 52). In business surveys, where the variables of interest are mainly metric, such as the amount of cash flow, heteroscedastic is often assumed. The choice $v_i = x_i$ results in the classical ratio estimator (cf. Breidt and Opsomer, 2017, p. 195). The assisting model ξ is used only to motivate the GREG estimator. Its unbiasedness does not depend on whether the population is really generated by the model. The efficiency is, indeed, influenced by the predictive power of the model ξ (cf. Särndal et al., 1992, p. 227, p. 239).

Definition 2. ***The Linear GREG Estimator***
The linear GREG estimator for the unknown population total $T_y = \sum_{i \in U} y_i$ relying on the linear regression model (2.3) is defined as

$$\hat{T}_y^{GREG} = \hat{T}_y^{HT} + \hat{\boldsymbol{B}}^T(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{HT}) \tag{2.4}$$

with

$$\hat{\boldsymbol{B}} = \left(\sum_{i \in s} \frac{\boldsymbol{x_i} \boldsymbol{x_i}^T}{\pi_i v_i} \right)^{-1} \sum_{i \in s} \frac{\boldsymbol{x}_i y_i}{\pi_i v_i} \tag{2.5}$$

as a design-consistent least squares estimate for the population parameter $\boldsymbol{\beta}$. It is assumed that the matrix $(\sum_{i \in s} \boldsymbol{x_i x_i}^T / \pi_i v_i)^{-1}$ is nonsingular.

It should be noted that if the matrix $(\sum_{i \in s} \boldsymbol{x_i x_i}^T / \pi_i v_i)^{-1}$ is singular, the generalized inverse can be applied for to invert the matrix.

In the following, we denote the GREG estimator given in Definition 2 as **naïve GREG estimator**. According to Definition 2, the GREG estimator can be interpreted as Horvitz-Thompson estimator expanded by an adjustment term. This adjustment term is composed of the difference between known and estimated totals of the auxiliaries weighted by the magnitude of the relationship between the variable of interest and the auxiliary variables. If the underlying model ξ has some predictive power, the adjustment term will often be negatively correlated with the error in the Horvitz-Thompson estimator, and therefore the GREG estimator is usually more precise than the Horvitz-Thompson estimator (cf. Särndal et al., 1992; Fuller, 2009).

The assisting models can have different forms. Firth and Bennett (1998) and Lehtonen and Veijanen (1998) first discussed non-linear assisting models. Firth and Bennett (1998) proposed canonical link generalized linear models and nonparametric models to include binary survey variables. Lehtonen and Veijanen (1998) considered multinomial logistic models to capture categorically distributed variables of interest. Local polynomial assisting models to derive a nonparametric GREG estimator are first considered by Breidt and Opsomer (2000). Montanari and Ranalli (2005) utilized nonparametric neural networks to assist the GREG estimator. Breidt and Opsomer (2009) examined nonparametric and semiparametric models to estimate densities and regression functions. Breidt et al. (2005) and McConville and Breidt (2013) discussed a penalized spline GREG estimator. Regression models with random components are considered by Park and Fuller (2009).

Remark 1. ***Categorical Variables as Auxiliaries***
In practice, the auxiliaries are often categorical variables, such as sex, age classes, marital status, or region. Consider a categorical variable with $p = 1, \dots, P$ mutually exclusive and exhaustive categories. Then, the P-vector $\boldsymbol{x_i} = (\gamma_{i1}, \dots, \gamma_{ip}, \dots, \gamma_{iP})^T$ with $\gamma_{ip} = 1$ if unit i belongs to category p of the auxiliary, and $\gamma_{ip} = 0$ otherwise specifies the category to which unit i belongs. An estimator only utilizing categorical variables as auxiliaries leads to the post-stratification estimator (cf. Holt and Smith, 1979; Valliant, 1993).

The GREG estimator can alternatively be expressed in linearly weighted form

$$\hat{T}_y^{\text{GREG}} = \sum_{i \in s} w_i y_i$$

with

$$w_i = \frac{1}{\pi_i} + \frac{\boldsymbol{x_i}^T}{\pi_i v_i}\left(\sum_{i\in s}\frac{\boldsymbol{x_i}\boldsymbol{x_i}^T}{\pi_i v_i}\right)^{-1}(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}). \tag{2.6}$$

The weights w_i depend only on the auxiliaries and inclusion probabilities. Once the auxiliaries are determined, the respective weights can be applied to any variable of interest. In contrast, non-linear or mixed-model GREG estimators require a separate model fitting for every single variable of interest. Hence, a global weight expression like (2.6) is not representable. An important property of the GREG estimator is that the sum of the weighted auxiliaries is consistent with the known population totals, that is $\sum_{i\in s} w_i\boldsymbol{x_i} = \boldsymbol{T_x}$. Särndal et al. (1992), Hidiroglou et al. (1995), Fuller (2002), Särndal (2007), and Kim and Park (2010) gave an comprehensive overview of the GREG estimator.

2.3.1 Approximate Design-Based Properties and Design-Based Variance

Because of the nonlinearity of the inverse within coefficient $\hat{\boldsymbol{B}}$ in (2.5), the GREG estimator is nonlinear. The nonlinearity prevents an analytical expression of the design-based variance being determined in a closed form. The Taylor linearization provides a remedy and approximates the non-linear GREG estimator by using Taylor series expansion. Based on the linearized version of the GREG estimator, an analytical expression of the design-based variance and design-based properties can be approved. We follow the derivation of the Taylor linearization given by Särndal et al. (1992, p. 173, p. 236). Rewriting the non-linear GREG estimator in (2.4) as function $f(\cdot)$ depending on different estimators yields

$$\begin{aligned}\hat{T}_y^{\text{GREG}} &= \hat{T}_y^{\text{HT}} + \hat{\boldsymbol{B}}^T(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})\\ &= f(\hat{T}_y^{\text{HT}}, \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}, \hat{\boldsymbol{D}}^{-1}, \hat{\boldsymbol{d}}),\end{aligned} \tag{2.7}$$

where the non-linear coefficient can be decomposed into $\hat{\boldsymbol{B}} = \hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{d}}$ with

$$\hat{\boldsymbol{D}} = \sum_{i\in s}\frac{\boldsymbol{x_i}\boldsymbol{x_i}^T}{\pi_i} \quad \text{as a } Q\times Q\text{-matrix with elements} \quad \hat{d}_{qq'} = \sum_{i\in s}\frac{x_{iq}x_{iq'}^T}{\pi_i} \quad \text{and}$$

$$\hat{\boldsymbol{d}} = \sum_{i\in s}\frac{\boldsymbol{x_i}y_i}{\pi_i} \quad \text{as a } Q\text{-vector with elements} \quad \hat{d}_{q0} = \sum_{i\in s}\frac{x_{iq}y_i}{\pi_i}.$$

A Taylor series of a real-valued function $f(\cdot)$, which is infinitely differentiable at a point a_0, is given by

$$f(a) = \sum_{n=0}^{\infty}\frac{f^{(n)}(a)}{n!}(a_0 - a)^n, \tag{2.8}$$

where $f^{(n)}(a)$ denotes the n-th derivative of $f(\cdot)$ evaluated at the point a. Thus, to approximate the non-linear function $f(\hat{T}_y^{\text{HT}}, \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}, \hat{\boldsymbol{D}}^{-1}, \hat{\boldsymbol{d}})$ by a linear function, we have to derive the first-order element of (2.8) and neglect the remainder term. To determine the partial derivatives of

$\hat{\boldsymbol{B}}$, we have to derive each element $\hat{d}_{qq'}$ and $\hat{d}_{q0}$ for $q = 1, \ldots, Q$ with respect to x_q. The partial derivatives of (2.7) are given by

$$\begin{aligned}
\frac{\partial f}{\partial \hat{T}_y^{\text{HT}}} &= 1, \\
\frac{\partial f}{\partial \hat{T}_{x_q}^{\text{HT}}} &= -B_q \qquad \text{for all } \; q = 1, \ldots, Q, \\
\frac{\partial f}{\partial \hat{d}_{qq'}} &= \Big(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}\Big)^T \Big(- \hat{\boldsymbol{D}}^{-1} \frac{\partial \hat{\boldsymbol{D}}}{\partial \hat{d}_{qq'}} \hat{\boldsymbol{D}}^{-1} \Big) \hat{\boldsymbol{d}} \\
&= \Big(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}\Big)^T \Big(- \hat{\boldsymbol{D}}^{-1} \boldsymbol{\Lambda_{qq'}} \hat{\boldsymbol{D}}^{-1} \Big) \hat{\boldsymbol{d}} \qquad \text{for all } \; q \leq q' = 1, \ldots, Q, \\
\frac{\partial f}{\partial \hat{d}_{q0}} &= (\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})^T \hat{\boldsymbol{D}}^{\boldsymbol{-1}} \boldsymbol{\Lambda_q} \qquad \text{for all } \; q = 1, \ldots, Q,
\end{aligned}$$

where $\boldsymbol{\Lambda_{qq'}}$ is a $Q \times Q$ matrix with the value 1 in positions (q, q') and (q', q), and 0 elsewhere; and $\boldsymbol{\Lambda_q}$ is a Q-vector with the value 1 in position q, and 0 elsewhere. Inserting these partial derivatives evaluated at the expected values $E(\hat{T}_y^{\text{HT}}) = T_y$, $E(\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\boldsymbol{HT}}) = \boldsymbol{T_x}$, $E(\hat{\boldsymbol{D}}) = \boldsymbol{D}$ and $E(\hat{\boldsymbol{d}}) = \boldsymbol{d}$ into equation (2.8), we obtain the first-order Taylor approximation given by

$$\begin{aligned}
\hat{T}_y^{\text{GREG}} \doteq \hat{T}_y^{\text{approx}} &= \Big(T_y + (\boldsymbol{T_x} - \boldsymbol{T_x})^T \boldsymbol{D^{-1} d} + 1 \cdot (\hat{T}_y^{\text{HT}} - T_y) + (-\sum_{q=1}^{Q} B_q (\hat{T}_{x_q} - T_{x_q})\Big) \\
&= \hat{T}_y^{\text{HT}} + \boldsymbol{B}^T (\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}). \qquad (2.9)
\end{aligned}$$

Accordingly, for large samples, when $\hat{T}_y^{\text{HT}}, \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}, \hat{\boldsymbol{D}}$, and $\hat{\boldsymbol{d}}$ take with high probability values close to $T_y, \boldsymbol{T_x}, \boldsymbol{D}$ and $\boldsymbol{d}$, the GREG estimator $\hat{T}_y^{\text{GREG}}$ will perform approximately as the linear estimator $\hat{T}_y^{\text{approx}}$ (cf. Särndal et al., 1992, p. 174). The estimators $\hat{T}_y^{\text{GREG}}$ and $\hat{T}_y^{\text{approx}}$ differ with respect to the true coefficient $\boldsymbol{B}$. Taylor linearization to approximate non-linear and complex statistics is well-established in the literature, as for example in Keyfitz (1957), Woodruff (1971), Demnati and Rao (2004), and Wolter (2007).

Based on the linearized estimator $\hat{T}_y^{\text{approx}}$, Särndal (1980) showed that the GREG estimator is approximately design-consistent. Cassel et al. (1976) proved the design-unbiasedness under mild design conditions for the assisting model and for the sampling design. Moreover, the design-based variance of $\hat{T}_y^{\text{approx}}$ provides a good approximation of the design-based variance of $\hat{T}_y^{\text{GREG}}$. The following definition originates from Särndal et al. (1992, p. 235).

Result 2. ***The Design-Based Variance of the GREG Estimator***
The approximate design-based variance of the GREG estimator, which is approximated through Taylor linearization, is given by

$$V(\hat{T}_y^{GREG}) = \sum_{i \in U} \sum_{j \in U} \triangle_{ij} \frac{R_i}{\pi_i} \frac{R_j}{\pi_j}$$

with residuals $R_i = y_i - \boldsymbol{x_i}^T\boldsymbol{B}$. *The variance estimator is obtained by*

$$\hat{V}(\hat{T}_y^{GREG}) = \sum_{i\in s}\sum_{j\in s}\frac{\triangle_{ij}}{\pi_{ij}}w_i r_i w_j r_j \tag{2.10}$$

with GREG weights w_i *defined by* (2.6) *and estimated residuals* $r_i = y_i - \boldsymbol{x_i}^T\hat{\boldsymbol{B}}$.

Proof. The linearized GREG estimator in (2.9) at the population level can be rewritten as

$$\begin{aligned}\hat{T}_y^{\text{approx}} &= \sum_{i\in U}\frac{y_i}{\pi_i} + \boldsymbol{B}^T\Big(\sum_{i\in U}\boldsymbol{x_i} - \sum_{i\in U}\frac{\boldsymbol{x_i}}{\pi_i}\Big)\\ &= \sum_{i\in U}\boldsymbol{x_i}^T\boldsymbol{B} + \sum_{i\in U}\Big(\frac{y_i - \boldsymbol{x_i}^T\boldsymbol{B}}{\pi_i}\Big)\\ &= \sum_{i\in U}\boldsymbol{x_i}^T\boldsymbol{B} + \sum_{i\in U}\frac{R_i}{\pi_i}.\end{aligned}$$

As the first term is constant, the approximated design-based variance of the GREG estimator is given by

$$\begin{aligned}V(\hat{T}_y^{\text{GREG}}) =& V(\hat{T}_y^{\text{approx}})\\ =& V\Big(\sum_{i\in U}\frac{R_i}{\pi_i}\Big)\\ =& \sum_{i\in U}\sum_{j\in U}\triangle_{ij}\frac{R_i}{\pi_i}\frac{R_j}{\pi_j}.\end{aligned}$$

$\hat{V}(\hat{T}_y^{\text{GREG}})$ results by estimating $V(\hat{T}_y^{\text{GREG}})$ from the sample s by the plug-in method. □

Therefore, the variance of the GREG estimator can be estimated via the variance of the residuals resulting from the assisting model ξ. Residuals indicate the distance between the predicted and the observed values. Therefore, we learn from Result 2 that even if the unbiasedness of the GREG estimator is independent of the correctness of the assisting model, its efficiency depends on the residuals obtained from the model. The theoretical justification for weighting the residuals in formula (2.10) with weights w_i defined in (2.6) instead of weighting with design weights d_i can be found in Särndal et al. (1989). Especially for small sample sizes, weighting with w_i reduces the bias of the variance estimator (cf. Deville et al., 1993). However, for small sample sizes, the Taylor linearization method generally has a tendency to underestimate the true variance (cf. Särndal et al., 1992, p. 176). Moreover, it should be noted that the variance estimation via Taylor linearization requires a separate formula for every variable of interest.

Alternatively, the variance of the GREG estimator can be obtained via resampling methods, such as jackknife, balanced repeated sampling or bootstrap methods. However, resampling methods are outside the scope of the thesis. The interested reader is referred to Wolter (2007) and Münnich (2008).

2.3.2 Asymptotic Properties

Asymptotic properties are obtained by increasing the sample size to infinity. Under certain conditions of the assumed model, Särndal (1980) and Wright (1983) showed that the GREG estimator is asymptotically design unbiased. The asymptotic design-consistency was proven by Isaki and Fuller (1982) and Robinson and Särndal (1983). The asymptotic design-unbiasedness and design-consistency, which are based on large sample properties, should be differentiated from approximate design-unbiasedness and design-consistency, which are based on Taylor linearization arguments.

2.3.3 Optimal GREG Estimator

Montanari (1987, p. 196) derived the optimal GREG estimator within the class of GREG estimators in the sense of minimizing the design-based variance. We will use the concept of the optimal GREG estimator in Chapters 5 and 6. Following Montanari (1987), the design-based variance of $\hat{T}_y^{\text{GREG}} = \hat{T}_y^{\text{HT}} + \boldsymbol{B}^T(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})$ is minimized by the coefficient

$$B^{\text{opt}} = [\text{V}(\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})]^{-1}\text{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}, \hat{T}_y^{\text{HT}}).$$

Because $\text{V}(\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})$ and $\text{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}, \hat{T}_y^{\text{HT}})$ are typically unknown, they have to be replaced by its consistent estimates $\widehat{\text{V}}(\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})$ and $\widehat{\text{Cov}}(\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}, \hat{T}_y^{\text{HT}})$ respectively, and we get

$$\hat{B}^{\text{opt}} = [\widehat{\text{V}}(\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})]^{-1}\widehat{\text{Cov}}(\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}, \hat{T}_y^{\text{HT}}). \tag{2.11}$$

The estimated coefficient $\hat{B}^{\text{opt}}$ is only asymptotically optimal (cf. Guandalini and Tillé, 2017, p. 3).

In practice, the estimation of B^{opt} might be intricate, because joint inclusion probabilities are required. As a remedy, for single-stage stratified sampling designs, Berger et al. (2003) proposed to include a stratification variable into the GREG estimator. Their proposed estimator of B^{opt} is easy to implement and relinquishes joint inclusion probabilities. Nangsue and Berger (2014) extended this estimator for two-stage samplings. Further discussion of the optimal GREG estimator can be found in Cochran (1977), Isaki and Fuller (1982), and Rao (1994).

2.3.4 GREG Estimators as Calibration Estimators

The GREG estimator can be seen as a special case of a broader class of calibration estimators, which will be important in Chapter 4. The class of calibration estimators has the form

$$\hat{T}_y^{\text{cal}} = \sum_{i \in s} w_i^{\text{cal}} y_i,$$

where the weights w_i^{cal} satisfy the calibration constraints

$$\sum_{i \in s} w_i^{\text{cal}} \boldsymbol{x_i} = \boldsymbol{T_x}. \tag{2.12}$$

Equation (2.12) guarantees that the sample sum of the weighted auxiliaries equals their known population totals.

Deville and Särndal (1992) first introduced the term *calibration*. In their **minimum distance approach**, the calibrated weights w_i^{cal} are chosen as close as possible to the original design weights d_i. Closeness between both weights is measured via a pre-specified distance function $G(w_i^{\text{cal}}, d_i)$. Requirements for the distance function are (i) $G(w_i^{\text{cal}}, d_i) \geq 0$; (ii) strict convexity; (iii) differentiability with respect to w_i^{cal} with $g(w_i^{\text{cal}}) = \frac{\partial G(w_i^{\text{cal}}, d_i)}{\partial w_i^{\text{cal}}}$, and (iv) $G(1) = g(1) = 0$ (cf. Haziza and Beaumont, 2017, p. 213). The latter property ensures that for $w_i^{\text{cal}} = d_i$ the distance is zero. Then the minimization problem is given by

$$\min_{w_i} \sum_{i \in s} \frac{d_i G(w_i^{\text{cal}}, d_i)}{\alpha_i} \quad \text{subject to calibration constraints (2.12),}$$

where α_i is a positive scale factor indicating the importance of unit i. The solution of the minimization problem yields the calibration weights

$$w_i^{\text{cal}} = d_i F(\alpha_i \boldsymbol{x_i}^T \boldsymbol{\lambda}), \tag{2.13}$$

where $F(u) = g^{-1}(u)$ is the inverse function of $g(\cdot)$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_Q)^T$ denotes a Q-vector of Lagrange multipliers. Properties (i) and (ii) ensure that the inverse function $F(u)$ exists. The Lagrange multiplier $\boldsymbol{\lambda}$ is determined by solving

$$\sum_{i \in s} d_i \boldsymbol{x_i} F(\alpha_i \boldsymbol{x_i}^T \boldsymbol{\lambda}) = \boldsymbol{T_x}. \tag{2.14}$$

As (2.14) involves a system of G equations and G unknowns, it can be solved via the Newton-Raphson algorithm (cf. Geiger and Kanzow, 2002, p. 235).

Applying the chi-square distance $G(w_i^{\text{cal}}, d_i) = \frac{1}{2}(\frac{w_i^{\text{cal}}}{d_i} - 1)^2$ and assuming $\alpha = 1$, we obtain

$$g(w_i^{\text{cal}}, d_i) = \left(\frac{w_i^{\text{cal}}}{d_i} - 1\right)\frac{1}{d_i} \quad \text{and} \quad F(\boldsymbol{x_i}^T \boldsymbol{\lambda}) = g^{-1}(\boldsymbol{x_i}^T \boldsymbol{\lambda}) = 1 + \boldsymbol{x_i}^T \boldsymbol{\lambda}.$$

Inserting $1 + \boldsymbol{x_i}^T \boldsymbol{\lambda}$ into (2.13) yields the calibrations weights

$$w_i^{\text{cal}} = d_i(1 + \alpha_i \boldsymbol{x_i}^T \boldsymbol{\lambda})$$

with Lagrange multipliers $\boldsymbol{\lambda} = (\sum_{i \in s} \alpha_i d_i \boldsymbol{x_i} \boldsymbol{x_i}^T)^{-1}(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})$. Hence, the minimization of the chi-square distance leads to the GREG weights defined in (2.6) with variance parameter $v_i = 1$ (cf. Särndal, 2007, p. 106). Calibration estimators associated with this distance function are also called generalized least squares (GLS) estimators, for example in Alexander (1987), Wu et al. (1997), Zieschang (1990), Verma and Clémenceau (1996), and Nieuwenbroek (1993).

Deville and Särndal (1992) examined six further distance functions, such as the Hellinger distance or the minimum entropy distance. Deville et al. (1993) introduced generalized raking estimators as a subclass of calibration estimators, which can be used when marginal counts of the auxiliaries are known. In this case, the distance function is multiplicative. The subclass of generalized raking estimators contains the classical raking estimator originated by Deming and Stephan (1940). Raking is equivalent to iterative proportional fitting and the maximum entropy approach by Wittenberg (2010) used in Chapter 3. Further distance functions are discussed in Huang and Fuller (1978), Alexander (1987), Singh and Mohl (1996), and Stukel et al. (1996).

Deville and Särndal (1992) showed that under mild conditions on $F(\cdot)$ the calibration estimator generated by different distance functions asymptotically equals the GREG estimator defined in (2.4). Thus, for large sample sizes, the choice of the distance function has only a minor impact on the properties of the calibration estimator. Singh and Mohl (1996) and Stukel et al. (1996) extended this finding to modest sample sizes.

An alternative derivation of calibration weights is obtained by the **functional form approach** considered by Estevao and Särndal (2000, 2006). Instead of a distance function, a simple functional form is imposed, which depends on $\boldsymbol{u_i} = (u_{i1}, \dots, u_{iQ})^T$. The vector $\boldsymbol{u_i}$ has to be of the same dimension as the auxiliary vector $\boldsymbol{x_i}$. Then, the calibrated weights w_i^{calF} are determined by the functional relationship

$$w_i^{\text{calF}} = d_i F(\boldsymbol{u_i}^T \boldsymbol{\lambda}^F), \tag{2.15}$$

where $\boldsymbol{\lambda}^F$ is a vector determined by the calibration constraints (2.12). F is a known real-valued function. Superscript F indicates functional form approach. For the linear function $F(z) = 1 + z$ the weights are given by $w_i^{\text{calF}} = d_i(1 + \boldsymbol{\lambda}^T \boldsymbol{u}_i)$ with $\boldsymbol{\lambda} = (\sum_{i \in s} d_i \boldsymbol{u}_i \boldsymbol{x}_i^T)^{-1}(\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x^{\text{HT}})$. The resulting calibration estimator is given by

$$\hat{\tau}_y^{\text{calF}} = \sum_{i \in s} w_i^{\text{calF}} y_i.$$

Inserting $\boldsymbol{\lambda}$ in w_i^{calF} gives

$$w_i^{\text{calF}} = d_i + d_i \boldsymbol{u}_i (\sum_{i \in s} d_i \boldsymbol{u}_i \boldsymbol{x}_i^T)^{-1} (\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x^{\text{HT}})$$

which reminds us of an instrumental variables regression known from econometrics. Therefore, the functional form approach was later termed as **instrument vector approach** by Estevao and Särndal (2006), Kott (2003) and Kott (2006). The vector $\boldsymbol{u_i}$ is supposed to be a function of the observed auxiliaries $\boldsymbol{x_i}$. The simple choice $\boldsymbol{u_i} = \alpha_i \boldsymbol{x_i}$ and $\alpha_i = v_i^{-1}$ yields the GREG weights defined in (2.6). The motivation of Estevao and Särndal (2000) behind the functional approach was that the change from the initial weight d_i to the calibrated weight w_i^{calF} can be controlled by appropriate choices of $\boldsymbol{u_i}$. Irrespective from the choice of u_i , the weights w_i^{calF} satisfy the calibration constraints.

Deville and Särndal (1992) and Kim and Park (2010) showed that the calibration estimator is design-consistent for T_y. Calibration estimators generated by different distance functions share the same large sample design-based variance (cf. Deville et al., 1993, p. 1014). Because Deville and Särndal (1992) proved the asymptotic equivalence of the calibration and the GREG estimator, the design-based variance of the calibration estimator can be approximated by the design-based variance defined in (2.10).

GREG and calibration estimators build on two different concepts. The GREG estimator is based on a linear relationship between the variable of interest and the auxiliaries. The calibration estimator indeed focuses more on the weights than on the assumption of an underlying regression model. A comprehensive discussion about the *GREG thinking* and the *calibration thinking* is given in Särndal (2007).

2.3.5 Avoiding Extreme Weights

Weights calculated according to (2.6), (2.13) or (2.15) can be very large or negative. The reasons for this might be small sample sizes or a variety of auxiliaries, or both. Weights reflect the probability of a unit to be sampled. Thus, negative weights or weights less than one can be interpreted as that the respective sampled unit does not even present itself, which is counter-intuitive. Nevertheless, it should be noted that negative weights do not influence the statistical properties of estimators. Large weights, in turn, can cause unstable estimations. Fortunately, considerable literature exists on methods to reduce the range of the weights.

Huang and Fuller (1978) first proposed a procedure that prevents extreme weights. Deville and Särndal (1992) and Deville et al. (1993) introduced some distance functions that produce weights that lie within a given range. Singh and Mohl (1996) compared several bounded distances by means of numerical examples. Husain (1969) and Isaki et al. (2004) used quadratic programming as an optimization method to set the weights boundary within a certain interval. Quadratic programming is equivalent to truncated linear calibration (cf. Park and Fuller, 2005, p. 8). Problems that arise due to bounding algorithms include slow convergence and multimodal weight distributions.

Théberge (2000) deduced conditions under which a solution of the optimization problem ensuring non-extreme weights exists. Tillé (1998) and Park and Fuller (2005) proposed a procedure that produces weights that are positive for the most samples. Chambers (1996) considered a ridge-type optimization problem under a certain coefficient matrix to produce non-negative weights.

Another possibility for avoiding extreme weights is to relax some of the calibration constraints. Rao and Singh (1997) studied a ridge shrinkage method for range-restricted weights, where the calibration constraints are satisfied within certain tolerances. Münnich et al. (2012b) developed a numeric algorithm that produces more stable and efficient solutions for calibration estimators, in particular applying box constraints. Münnich et al. (2012b) reformulated the calibration problem as a constrained optimization problem and rewrote it as a nonlinear system of non-differentiable equations. It is given by

$$\min_{w_i} \sum_{i \in s} \frac{(w_i - d_i)^2}{d_i}$$

subject to

$$\begin{aligned} \sum_{i \in s} w_i \boldsymbol{x_i} &= \boldsymbol{T_x} \\ L \leq & w_i \leq U \end{aligned}$$

with L as lower bound and U as upper bound. The minimization problem was solved via a Lagrangian approach using a highly efficient semi-smooth Newton method within a moderate computing time even for problems of high dimensions.

2.4 Cluster Sampling

Household surveys are often realized by means of cluster sampling, where the finite population is divided into subpopulations of units, called clusters. Potential subpopulations are geographical districts, city blocks, enterprises, establishments, schools, or some type of aggregate unit. Cluster sampling is often implemented, if no complete, up-to-date and accessible list containing all population units that can serve as sampling frame is available (cf. Hansen et al., 1953, p. 243). A further reason to implement cluster sampling is that the cost of selecting clusters might be negligible compared to the cost of selecting individual units (cf. Lohr, 2009, p. 170). That may be the case if the data are collected through personal interviews. Dividing the population into clusters, for example areas, can reduce travel costs of interviewers (cf. Valliant et al., 2013, p. 203).

One can distinguish between two-stage cluster sampling and single-stage cluster sampling. **Single-stage cluster sampling** is characterized by a complete enumeration within a selected cluster. This implies that all units within a cluster are sampled. In contrast, in **two-stage cluster sampling**, only a subsample of units is selected within a cluster. The focus of this thesis lies on single-stage cluster sampling, because in household surveys, often all persons within a drawn household are selected into the sample.

To formalize the sampling process of single-stage cluster sampling, we follow Särndal et al. (1992, p. 127). A finite population $U_p = \{1, \ldots, i, \ldots, N\}$ of persons can be partitioned into subpopulations called primary sampling units (PSU). In a household survey, the PSU are households. The sampling process of household surveys involves two stages:

1) At the first stage, from a finite population of households $U_h = \{1, \dots, g, \dots, M\}$ a sample s_h is selected according to the sampling design $p(\cdot)$, where $p(s_h)$ is the probability of selecting s_h. The sample size of s_h is m. Let U_g be the population of persons within household g of size N_g. The probability sampling design generates for every household g a known inclusion probability $\pi_g = Pr(g \in s_h) = \sum_{s_h : g \in s_h} p(s_h)$ with $\pi_g > 0$.

2) At the second stage, all persons within a selected household, called secondary sampling units (SSU), are sampled. The finite population and the sample of persons are denoted by $U_p = \cup_{g \in U_h} U_g$ and $s_p = \cup_{g \in s_h} U_g$, respectively. The sample size is given by $n = \sum_{g \in s_h} N_g$. The first-order inclusion probability of person i induced by the design $p(\cdot)$ is given by

$$\pi_i = Pr(i \in s_p) = Pr(g \in s_h) = \pi_g.$$

All first-order inclusion probabilities are equal for all $i \in U_g$. The second-order inclusion probabilities are given by

$$\pi_{ij} = Pr(i, j \in s_p) = Pr(g \in s_h) = \pi_g$$

if both i and j belong to the same household g, and

$$\pi_{ij} = Pr(i, j \in s_p) = Pr(g \,\&\, k \in s_h) = \pi_{gk}$$

if i and j belong to different households g and k. Note that $\pi_{ii} = \pi_i$.

The efficiency of cluster sampling depends on the internal composition of the clusters. The more homogeneous the clusters are, the less efficient the cluster sampling. In practice, many naturally formed clusters are characterized to be rather homogeneous with small within-cluster variation (cf. Lehtonen and Pahkinen, 2004, p. 83). General differences between cluster sampling and simple random sampling are revealed by Hansen et al. (1953, pp. 259-270). A comparison between cluster sampling and stratified sampling can be found in Lohr (2009, p. 167).

To clarify the difference between single-stage cluster sampling and simple random sampling, the following result outlines the Horvitz-Thompson estimator under single-stage cluster sampling, which is directly comparable with Result 1. Note that $y_g = \sum_{i \in U_g} y_i$.

Result 3. ***The Horvitz-Thompson Estimator under Single-Stage Cluster Sampling***
Under single-stage cluster sampling, the Horvitz-Thompson estimator of the population total $T = \sum_{g \in U_h} y_g$ is given by

$$\hat{T}_y^{HT} = \sum_{g \in s_h} \frac{y_g}{\pi_g}$$

with variance

$$V(\hat{T}_y^{HT}) = \sum_{g \in U_h} \sum_{k \in U_h} \triangle_{gk} \frac{y_g}{\pi_g} \frac{y_k}{\pi_k},$$

where $\triangle_{gk} = \pi_{gk} - \pi_g \pi_k$*. Given that* $\pi_{gk} > 0$ *for all* $g, k \in U_h$*, an unbiased estimator of* $V(\hat{T}_y^{HT})$ *is given by*

$$\hat{V}(\hat{T}_y^{HT}) = \sum_{g \in s_h} \sum_{k \in s_h} \frac{\triangle_{gk}}{\pi_{gk}} \frac{y_g}{\pi_g} \frac{y_k}{\pi_k}. \tag{2.16}$$

Proof. See Särndal et al. (1992, p. 128). □

Comparing Results 1 and 3, it becomes obvious that the formulas are derived at the aggregated cluster level in the latter case. The formulas in Result 3 are valid for various designs to sample the households at the first stage. If the households are selected via simple random sampling, the variance estimator in (2.16) simplifies to

$$\hat{V}_{SSCS}(\hat{T}_y^{\text{HT}}) = \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \frac{1}{m-1} \sum_{g \in s_h} (y_g - \bar{y}^h)^2$$

with $\bar{y}^h = m^{-1} \sum_{g \in s_h} y_g$ as mean value. Superscript h emphasizes the difference to the mean value at the person level $\bar{y} = n^{-1} \sum_{i \in s} y_i$ used in (2.2). This sampling design is denoted as simple single-stage cluster sampling (SSCS).

3 Integrated Weighting

Household surveys provide information about both person and household characteristics. When estimating the same characteristic based on either a person- or a household-level data set, the question arises of how to ensure consistency between both estimates. For example, the estimated total of household income should coincide with the total income estimated at the person level. The current practice of statistical offices to ensure consistent estimates in household surveys is integrated weighting originated by Lemaître and Dufour (1987). The method of integrated weighting produces one single weight for all persons within the same household by substituting the original auxiliary information at the person level by the corresponding household mean values. This single integrated person weight is then assigned one-to-one to the household to which the person belongs. Instead of calculating integrated person weights, Nieuwenbroek (1993) proposed to calculate integrated household weights based on aggregated auxiliary information. This integrated household weight is then assigned one-to-one to all persons within the same household. Consistency is thereby ensured by the same weights used to estimate person- and household-level characteristics. The use of the same weights for all persons within the same household is supported by several authors:

- Lemaître and Dufour (1987, p. 199): "[...] a[n] [integrated] method [...] would be appropriate for both individual and family estimation.“

- Nieuwenbroek (1993, p. 6): "[...] weighting methods that give one weight per household are relevant."

- Lavallée (1995, p. 27): "Note that the fact of allocating the same weight to all units has the considerable advantage of ensuring consistency of estimates for units and clusters."

- Estevao and Särndal (2006, p. 139): "It is practical to give all units within a cluster the same weight in computing unit statistics, and use this weight for computing cluster statistics."

- Verma et al. (2006, p. 10): "It is desirable, therefore, to use a weighting procedure which ensures consistency between analyses involving the two types of units. The recommended procedure is integrative weighting [...]."

- Särndal (2007, p. 113): "Integrated weighting is often used in practice."

- Steel and Clark (2007, p. 51): "It is sometimes convenient to have equal weights for people within a household, for surveys which collect information on both household and person level variables of interest."

- Branson and Wittenberg (2014, p. 20): “Given that the household is the unit that is sampled, it makes more sense for person weights to be common within a household.”

- van den Brakel (2016, p. 149): “[...] it is relevant to apply a weighting method which yields one unique household weight for all its members [...].”

For example, integrated weighting is currently implemented in the German Microcensus (cf. Afentakis and Bihler, 2005), the Britain Integrated Household Survey (cf. ONS, 2012), the Finnish Time Use Survey (cf. Väisänen, 2002), the Canadian Labor Force Survey (LFS) (cf. Statistics Canada, 2017), and in the Swiss Household Panel (cf. Antal and Rothenbühler, 2015). It is also recommended by Eurostat for EU-SILC (cf. European Commission, 2014, p. 37).

The support for integrative weighting seems surprising given that equal weights for all persons within a household and the household itself is a very strict requirement. Equal weights no longer reflect the heterogeneity of the individual persons within a household and the individual patterns of the persons are lost. It is intuitive that for very volatile variables, such as income, the resulting estimates might be significantly influenced when the same weights are assigned to all persons within a household, independently of whether they are top earners, children, or inactive persons. Therefore, this chapter aims to answer the following question: What are the consequences of the strict restriction of equal weights for the estimation of person characteristics? We start by reviewing the theory on integrated weighting in Section 3.1. Section 3.2 adduces potential consequences of integrated weighting due to the enforcement of equal weights. Section 3.3 reviews the empirical evidence in the literature. A simulation study (Section 3.4) evaluates the aforementioned consequences by comparing the performance of an integrated and a naïve GREG estimator. Section 3.5 concludes with a summary.

3.1 Theory of Integrated Weighting

This section reviews the literature on integrated weighting as current practice in statistical offices. After unveiling a considerably important property of integrated weighting (Section 3.1.1), which is so far to the best of our knowledge neglected in the literature, we discuss two different approaches of integrated weighting based on persons (Section 3.1.2) and based on households (Section 3.1.3). In Section 3.1.4, we combine both approaches to one more general approach. The generalization facilitates a comparison of both approaches in Section 3.1.5. Section 3.1.6 presents a different concept of integrated weighting.

3.1.1 The Integrated Property

In the integrated weighting approach, consistency between person- and household-level estimates is ensured by calculating weights at one level and then assigning these weights one-to-one to the other level. As a consequence thereof, it is not necessarily guaranteed that

- the weights at the person level sum up to the number of persons in the population and simultaneously
- the weights at the household level sum up to the number of households in the population.

We define the compliance to both points as the **integrated property**. An additional variable has to be incorporated into the auxiliary variables to ensure the integrated property. Therefore, we define

$$\boldsymbol{x}_{\boldsymbol{i}}^{\circ} = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{iQ})^T = (N_g^{-1},\ 1, x_{i2}, \ldots, x_{iQ})^T = (N_g^{-1}, \boldsymbol{x}_{\boldsymbol{i}})^T$$

as the integrated auxiliary vector of person i of dimension $(Q+1)$, which sums up within each household g to

$$\boldsymbol{x}_{\boldsymbol{g}}^{\circ} = (x_{g0}, x_{g1}, x_{g2}, \ldots, x_{gQ})^T = (\ 1,\ N_g,\ x_{g2}, \ldots, x_{gQ})^T$$

as the integrated auxiliary vector of household g of dimension $(Q+1)$. Superscript $\circ$ indicates the integrated property. The additional auxiliary variable enforcing the integrated property at the person level is $x_{i0} = N_g^{-1}$, which sums to one per household. At the household level, the additional auxiliary is given by $x_{g1} = N_g$, whose person-level counterpart is the intercept. We differentiate between a person-level intercept, $x_{i1} = 1$, and a household-level intercept, $x_{g0} = 1$. It is important to note that the integrated property is required only in the case of integrated weighting where the weights are assigned one-to-one between the levels. For analytical and programming purposes, the order of the auxiliaries is crucial. To the best of our knowledge, the integrated property has so far been neglected, or at least not explicitly mentioned, in the literature.

The corresponding known and estimated total vector of dimension $(Q+1)$ are denoted by $\boldsymbol{T}_{x^{\circ}} = (M, \boldsymbol{T}_x)^T$ and $\hat{\boldsymbol{T}}_{x^{\circ}}^{\text{HT}} = (\hat{T}_{x_0}^{\text{HT}}, \hat{\boldsymbol{T}}_x^{\text{HT}})^T$, respectively.

3.1.2 Integrated GREG Estimator with Persons as Basis

Before integrated weighting was introduced, a widely used method to produce weights for household surveys was the **principal person method** offered by Alexander (1987). According to this method, the design weights at the person level were adjusted via post-stratification to meet known auxiliary totals. Because the individual auxiliaries differ from person to person, the resulting weights also differ within a household. Household weights are determined by the weight of one household member, the principal person. For example, in the Canadian LFS and the U.S. Consumer Expenditure Survey, this principal person was the female spouse unless there was a single male head. The choice of a female principal person was justified, as women tend to have a better coverage rate than men (cf. Zieschang, 1990, p. 985). The main disadvantage of the principal person method is that regardless of who is declared as principal person, consistency between person- and household-level estimates is not guaranteed. Moreover, the household decomposition is not taken into account. A comprehensive overview of the principal person method can be found in Hanson (1978, Chapter 5) and in Alexander (1987).

To overcome these disadvantages, Lemaître and Dufour (1987) first introduced integrated weighting. Their proposed integrated GREG estimator produces one single weight for all persons within a household. The integrated weights are appropriate for both person- and household-level estimation. In the following, integrated weighting and integrated GREG estimator will be used interchangeably. To produce equal weights, the original auxiliaries $\boldsymbol{x}_i^\circ$ are replaced by the corresponding household mean values. The household mean value is determined by $\bar{\boldsymbol{x}}_g^\circ = N_g^{-1} \sum_{i \in U_g} \boldsymbol{x}_i^\circ$ which is assigned to all persons within the household. Then, at the person level the household mean value is denoted as $\bar{\boldsymbol{x}}_i^\circ$. It should be noted that $\bar{\boldsymbol{x}}_i^\circ$ is the mean value per household, not the mean value of all sampled persons. We choose the subscript i to emphasize that $\bar{\boldsymbol{x}}_i^\circ$ is a person-level variable, even if it has the same value for all persons within the same household. Replacing $\bar{\boldsymbol{x}}_i^\circ$ as auxiliaries into the assisting linear regression model ξ yields

$$y_i = \bar{\boldsymbol{x}}_i^{\circ T} \boldsymbol{\beta_p} + \epsilon_i \quad \text{for all} \quad i \in U_p \tag{3.1}$$

with $\boldsymbol{\beta_p}$ as population regression coefficient at the person level and ϵ_i as unobserved random error. Note that $E_\xi(\epsilon_i) = 0$, $V_\xi(\epsilon_i) = v_i \sigma^2$, and $\mathrm{E}_\xi(\epsilon_i \epsilon_j) = 0$ for all $i \neq j$. Lemaître and Dufour (1987) assumed a constant variance parameter $v_i = 1$ for all $i \in U_p$. This is equivalent to assuming homoscedasticity. The assisting linear regression model (3.1) results in the integrated person-level GREG estimator for the unknown total $T_{y_p} = \sum_{i \in U_p} y_i$

$$\hat{T}_{y_p}^{\mathrm{LD}} = \hat{T}_{y_p}^{\mathrm{HT}} + \hat{\boldsymbol{B}}_{\boldsymbol{p}}^{\mathrm{LD}^T} (\boldsymbol{T}_{\boldsymbol{x}^\circ} - \hat{\boldsymbol{T}}_{\boldsymbol{x}^\circ}^{\mathrm{HT}}) \tag{3.2}$$

with

$$\hat{\boldsymbol{B}}_{\boldsymbol{p}}^{\mathrm{LD}} = \left(\sum_{i \in s_p} \frac{\bar{\boldsymbol{x}}_i^\circ \bar{\boldsymbol{x}}_i^{\circ T}}{\pi_i} \right)^{-1} \sum_{i \in s_p} \frac{\bar{\boldsymbol{x}}_i^\circ y_i}{\pi_i}$$

as vector containing $(Q+1)$ coefficients. The superscript LD refers to Lemaître and Dufour. It is assumed that the matrix $(\sum_{i \in s_p} \bar{\boldsymbol{x}}_i^\circ \bar{\boldsymbol{x}}_i^{\circ T} / \pi_i)^{-1}$ is nonsingular. $\boldsymbol{T}_{\boldsymbol{x}^\circ}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{x}^\circ}^{\mathrm{HT}}$ are the known and estimated $(Q+1)$-vectors of the integrated auxiliary totals, respectively. The integrated person weights generated by (3.2) are given by

$$w_i^{\mathrm{LD}} = \frac{1}{\pi_i} + \frac{\bar{\boldsymbol{x}}_i^{\circ T}}{\pi_i} \left(\sum_{i \in s_p} \frac{\bar{\boldsymbol{x}}_i^\circ \bar{\boldsymbol{x}}_i^{\circ T}}{\pi_i} \right)^{-1} (\boldsymbol{T}_{\boldsymbol{x}^\circ} - \hat{\boldsymbol{T}}_{\boldsymbol{x}^\circ}^{\mathrm{HT}}). \tag{3.3}$$

Note that under single-stage cluster sampling, $\pi_i = \pi_j$ for all $i, j \in U_g$. The weights (3.3) are equal for all persons with the same household, because the household members share the same auxiliary vector $\bar{\boldsymbol{x}}_i^\circ$. The person weight is assigned one-to-one to the corresponding household; consequently, $w_g^{\mathrm{LD}} = w_i^{\mathrm{LD}}$ for all $i \in U_g$. Hence, consistency is ensured by assigning the same weight to all persons within a household and to the household itself. Given that $\sum_{i \in U_p} \bar{\boldsymbol{x}}_i^\circ = \boldsymbol{T}_{\boldsymbol{x}^\circ} = \sum_{g \in U_h} \boldsymbol{x}_g^\circ$, it is easy to verify that $\sum_{i \in s_p} w_i^{\mathrm{LD}} \bar{\boldsymbol{x}}_i^\circ = \boldsymbol{T}_{\boldsymbol{x}^\circ}$ and $\sum_{g \in s_h} w_g^{\mathrm{LD}} \boldsymbol{x}_g^\circ = \boldsymbol{T}_{\boldsymbol{x}^\circ}$. That is, the sample sums of the weighted auxiliaries meet the known totals at both levels.

As the integrated GREG estimator is a special case of the GREG estimator as defined in Section 2.3, it is also asymptotically unbiased. The variance estimator under single-stage cluster sampling approximated by Taylor linearization can be obtained by

$$\hat{V}(\hat{T}_{y_p}^{\mathrm{LD}}) = \sum_{g \in s_h} \sum_{k \in s_h} \frac{\triangle_{gk}}{\pi_{gk}} w_g^{\mathrm{LD}} r_g^{\mathrm{LD}} w_k^{\mathrm{LD}} r_k^{\mathrm{LD}}$$

with $\triangle_{gk} = \pi_{gk} - \pi_g \pi_k$ and w_g^{LD} defined in (3.3). The estimated residuals are determined by $r_g^{\mathrm{LD}} = y_g - \bar{\boldsymbol{x}}_{\boldsymbol{g}}^{\circ T} \hat{\boldsymbol{B}}_{\boldsymbol{p}}^{\mathrm{LD}}$.

The integrated GREG estimator differs from the naïve GREG estimator in two respects. Firstly, constructed household mean values are utilized instead of the original auxiliaries. Secondly, the integrated GREG estimator requires an additional auxiliary variable, N_g^{-1}, to ensure the integrated property (see Section 3.1.1).

Alternatively to the integrated GREG estimator (3.2) with persons as basis, Branson and Wittenberg (2014) suggested producing integrated person weights using minimum cross-entropy estimator. The cross-entropy estimation approach is based on arguments of information theory (cf. Golan et al., 1997). It attempts to minimize the information loss from moving a prior weight distribution to a post-calibration distribution (cf. Branson and Wittenberg, 2014, p. 26). The information loss is minimized subject to the linear constraints that the final weights (i) meet the known totals, (ii) be close as possible to the initial weights, and (iii) be equal for all household members. The last constraint ensures consistency between person- and household-level estimates. Wittenberg (2010) showed that the minimum cross-entropy approach is equivalent to raking introduced by Deming and Stephan (1940) and to a calibration estimator with a multiplicative distance function introduced by Deville and Särndal (1992) and Deville et al. (1993). Compared to the weights suggested by Lemaître and Dufour (1987) generated by formula 3.3, the minimum cross-entropy weights are prevented from being negative. However, extreme weights might occur.

3.1.3 Integrated GREG Estimator with Households as Basis

Nieuwenbroek (1993) proposed calculating the integrated weights at the household level. Suppose $\boldsymbol{x}_{\boldsymbol{g}}^{\circ} = \sum_{i \in U_g} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^{\circ}$ is the per-household aggregated person-level information. The assumed linear regression model ξ relating the variable of interest and the auxiliaries at household level is given by

$$y_g = \boldsymbol{x}_{\boldsymbol{g}}^{\circ T} \boldsymbol{\beta}_{\boldsymbol{h}} + \epsilon_g \quad \text{for all} \quad g \in U_h \tag{3.4}$$

with $\boldsymbol{\beta}_{\boldsymbol{h}}$ as population regression coefficient at the household level and ϵ_g as unobserved random error. Note that $E_\xi(\epsilon_g) = 0$, $V_\xi(\epsilon_g) = v_g \sigma^2$, and $\mathrm{E}_\xi(\epsilon_g \epsilon_k) = 0$ for all $g \neq k$. To estimate household-level variables of interest, Nieuwenbroek (1993) suggested to set the variance component v_g proportional to the household size. For variables not correlated with the household size he suggested $v_g = 1$. The same recommendations for the variance components can be found in van den Brakel (2013, 2016). As we are interested in the estimation of person characteristics, we assume $v_g = 1$ in the following. The assisting model (3.4) results in the integrated GREG estimator for the unknown population household-level total $T_{y_h} = \sum_{g \in U_h} y_g$

$$\hat{T}_{y_h}^{\mathrm{N}} = \hat{T}_{y_h}^{\mathrm{HT}} + \hat{\boldsymbol{B}}_{\boldsymbol{h}}^{\mathrm{N}^T} (\boldsymbol{T}_{\boldsymbol{x}^\circ} - \hat{\boldsymbol{T}}_{\boldsymbol{x}^\circ}^{\mathrm{HT}}) \tag{3.5}$$

with

$$\hat{\boldsymbol{B}}_{\boldsymbol{h}}^{\mathrm{N}} = \left(\sum_{g \in s_h} \frac{\boldsymbol{x}_{\boldsymbol{g}}^{\circ} \boldsymbol{x}_{\boldsymbol{g}}^{\circ T}}{\pi_g} \right)^{-1} \sum_{g \in s_h} \frac{\boldsymbol{x}_{\boldsymbol{g}}^{\circ} y_g}{\pi_g}. \tag{3.6}$$

It is assumed that the matrix $(\sum_{g \in s_h} \boldsymbol{x}_{\boldsymbol{g}}^{\circ} \boldsymbol{x}_{\boldsymbol{g}}^{\circ T} / \pi_g)^{-1}$ is nonsingular. The superscript N refers to Nieuwenbroek. It should be remarked that since the integrated GREG estimator uses the same auxiliaries at the person and household level, the vectors $\boldsymbol{T}_{\boldsymbol{x}^\circ}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{x}^\circ}^{\mathrm{HT}}$ are valid at both levels.

The integrated household-level weights generated by (3.5) are given by

$$w_g^{\mathrm{N}} = \frac{1}{\pi_g} + \frac{\boldsymbol{x}_{\boldsymbol{g}}^{\circ T}}{\pi_g} \left(\sum_{g \in s_h} \frac{\boldsymbol{x}_{\boldsymbol{g}}^{\circ} \boldsymbol{x}_{\boldsymbol{g}}^{\circ T}}{\pi_g} \right)^{-1} (\boldsymbol{T}_{\boldsymbol{x}^\circ} - \hat{\boldsymbol{T}}_{\boldsymbol{x}^\circ}^{\mathrm{HT}}). \tag{3.7}$$

The household weight w_g^N is then assigned one-to-one to all persons within the same household; consequently, $w_i^{\mathrm{N}} = w_g^{\mathrm{N}}$ for all $i \in U_g$. It is easy to verify that $\sum_{g \in s_h} w_g^{\mathrm{N}} \boldsymbol{x}_{\boldsymbol{g}}^{\circ} = \boldsymbol{T}_{\boldsymbol{x}^\circ} = \sum_{i \in s_p} w_i^{\mathrm{N}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^{\circ}$. This implies that consistent estimates at both levels are guaranteed.

Nieuwenbroek (1993) proved that for $v_g = N_g$, the integrated weights with persons as basis defined in (3.3) and the integrated weights with households as basis defined in (3.7) are equivalent. However, Nieuwenbroek (1993) neglected that the equality of both integrated weights is valid if and only if the auxiliary vector at the person level contains N_g^{-1} and the auxiliary vector at the household level contains N_g as additional variables. Therefore, the derived equivalence of both approaches is valid only if the integrated property is fulfilled.

Analogously to its person-level counterpart, the integrated GREG estimator at the household level is asymptotically unbiased. The variance estimator under single-stage cluster sampling approximated by Taylor linearization can be obtained by the residual variance

$$\hat{V}(\hat{T}_{y_h}^{\mathrm{N}}) = \sum_{g \in s_h} \sum_{k \in s_h} \frac{\triangle_{gk}}{\pi_{gk}} w_g^{\mathrm{N}} r_g^{\mathrm{N}} w_k^{\mathrm{N}} r_k^{\mathrm{N}}$$

with estimated residuals $r_g^{\mathrm{N}} = y_g - \boldsymbol{x}_{\boldsymbol{g}}^{\circ T} \hat{\boldsymbol{B}}_{\boldsymbol{h}}^{\mathrm{N}}$ and w_g^{N} as defined in (3.7).

As an alternative to the integrated GREG estimator (3.5) with households as basis, Zieschang (1986), Luery (1986), and Alexander (1987) discussed a GLS approach to produce household weights by minimizing the distance between the initial design weights and the resulting calibration weights. When these weights are applied for both person and household characteristics, the GLS approach is asymptotically equivalent to (3.5). Verma and Clémenceau (1996) suggested extending the household auxiliaries in the GLS approach to include person information. For this purpose, the sample distribution of the person auxiliaries is inflated by the household size. The weights produced by the extended GLS estimator are asymptotically equivalent to the integrated household weights generated by (3.7) with additional person-level information.

Isaki et al. (2004) used quadratic programming to produce household weights suitable for the estimation of person and household characteristics. Quadratic programming seeks household weights that minimize a quadratic objective function subject to the linear constraints that the final weights (i) are as close as possible to the initial weights, (ii) are within certain bounds, (iii) meet known person- and household-level totals, and (iv) are design-consistent. Park and Fuller (2005, p. 8) showed that quadratic programming is equal to the truncated linear distance function introduced by Deville and Särndal (1992). When dropping the bounds in constraint (ii), quadratic programming generates weights that are asymptotically equivalent to the weights (3.7) suggested by Nieuwenbroek (1993).

The (extended) GLS calibration estimator and quadratic programming with an integrated GREG estimator with households as basis introduced in this section show equivalence. Thus, we do not pursue the different approaches separately in the following.

3.1.4 Combining Both Integrated GREG Estimators into One Single Estimator

The weights are calculated at one level and then assigned one-to-one to the other level for both integrated GREG estimators introduced in Sections 3.1.2 and 3.1.3. These weights are used for person- and household-level estimation. Note that with $\boldsymbol{x}_g^\circ = N_g \bar{\boldsymbol{x}}_i^\circ$ for $i \in U_g$, $\pi_i = \pi_g$ and $\sum_{i \in s_p} = \sum_{g \in s_h} \sum_{i \in U_g}$, the integrated household-level coefficient (3.6) can be rewritten as

$$\begin{aligned}
\hat{\boldsymbol{B}}_h^{\mathrm{N}} &= \left(\sum_{g \in s_h} \frac{\boldsymbol{x}_g^\circ \boldsymbol{x}_g^{\circ T}}{\pi_g} \right)^{-1} \sum_{g \in s_h} \frac{\boldsymbol{x}_g^\circ y_g}{\pi_g} \\
&= \left(\sum_{g \in s_h} \frac{N_g \bar{\boldsymbol{x}}_i^\circ N_g \bar{\boldsymbol{x}}_i^{\circ T}}{\pi_i} \right)^{-1} \sum_{g \in s_h} \frac{N_g \bar{\boldsymbol{x}}_i^\circ y_g}{\pi_i} \\
&= \left(\sum_{g \in s_h} \sum_{i \in U_g} \frac{N_g \bar{\boldsymbol{x}}_i^\circ N_g \bar{\boldsymbol{x}}_i^{\circ T}}{N_g \pi_i} \right)^{-1} \sum_{g \in s_h} \sum_{i \in U_g} \frac{N_g \bar{\boldsymbol{x}}_i^\circ y_i}{\pi_i} \\
&= \left(\sum_{i \in s_p} \frac{N_g \bar{\boldsymbol{x}}_i^\circ \bar{\boldsymbol{x}}_i^{\circ T}}{\pi_i} \right)^{-1} \sum_{i \in s_p} \frac{N_g \bar{\boldsymbol{x}}_i^\circ y_i}{\pi_i} \qquad (3.8)
\end{aligned}$$

given $\sum_{g \in U_h} \boldsymbol{x}_g^\circ = \sum_{g \in U_h} \sum_{i \in U_g} \bar{\boldsymbol{x}}_i^\circ / N_g$. According to (3.8), the coefficient $\hat{\boldsymbol{B}}_h^{\mathrm{N}}$ can either be calculated at the household level based on the assisting model (3.4), or it can equivalently be calculated at the person level under the assisting model (3.1) with variance parameter $v_i = N_g^{-1}$.

Therefore, we generalize both integrated GREG estimators defined in (3.2) and (3.5) to one single estimator calculated at the person level but with different variance parameters.

Definition 3. ***The Integrated GREG Estimator***
The integrated GREG estimator relying on the assisting person-level model (3.1) *can be expressed by*

$$\hat{T}^{INT}_{y_p} = \hat{T}^{HT}_{y_p} + \hat{\boldsymbol{B}}^{\circ T}_{\boldsymbol{p}} (\boldsymbol{T}_{\boldsymbol{x}^\circ} - \hat{\boldsymbol{T}}^{\boldsymbol{HT}}_{\boldsymbol{x}^\circ}), \tag{3.9}$$

where

$$\hat{\boldsymbol{B}}^\circ_{\boldsymbol{p}} = \left(\sum_{i\in s_p} \frac{\bar{\boldsymbol{x}}^\circ_i \bar{\boldsymbol{x}}^{\circ T}_i}{\pi_i v_i}\right)^{-1} \sum_{i\in s_p} \frac{\bar{\boldsymbol{x}}^\circ_i y_i}{\pi_i v_i} \tag{3.10}$$

is a vector containing $(Q+1)$ *person-level coefficients. The corresponding integrated person weights are given by*

$$w^{INT}_i = \frac{1}{\pi_i} + \sum_{i\in s_p} \frac{\bar{\boldsymbol{x}}^{\circ T}_i}{\pi_i v_i} \left(\sum_{i\in s_p} \frac{\bar{\boldsymbol{x}}^\circ_i \bar{\boldsymbol{x}}^{\circ T}_i}{\pi_i v_i}\right)^{-1} (\boldsymbol{T}_{\boldsymbol{x}^\circ} - \hat{\boldsymbol{T}}^{\boldsymbol{HT}}_{\boldsymbol{x}^\circ})$$

which are equivalent to the household weights

$$\begin{aligned} &= \frac{1}{\pi_g} + \sum_{g\in s_h} \frac{\boldsymbol{x}^{\circ T}_g}{\pi_g v_g} \left(\sum_{g\in s_h} \frac{\boldsymbol{x}^\circ_g \boldsymbol{x}^{\circ T}_g}{\pi_g v_g}\right)^{-1} (\boldsymbol{T}_{\boldsymbol{x}^\circ} - \hat{\boldsymbol{T}}^{\boldsymbol{HT}}_{\boldsymbol{x}^\circ}) \\ &= w^{INT}_g \end{aligned}$$

for all $i \in U_g$. *Note that* $v_g = \sum_{i\in U_g} v_i$. *The variance estimator can be approximated by Taylor linearization*

$$\hat{V}(\hat{T}^{INT}_{y_p}) = \sum_{g\in s_h} \sum_{k\in s_h} \frac{\triangle_{gk}}{\pi_{gk}} w^{INT}_g r^{INT}_g w^{INT}_k r^{INT}_k \tag{3.11}$$

with residuals $r^{INT}_g = y_g - \boldsymbol{x}^{\circ T}_g \hat{\boldsymbol{B}}^\circ_{\boldsymbol{p}}$ *and* $\triangle_{gk} = \pi_{gk} - \pi_g \pi_k$.

According to Definition 3, the introduced integrated GREG estimators based on persons (see Section 3.1.2) and based on households (see Section 3.1.3) differ with respect to the variance component. Inserting a variance component of $v_i = 1$ into $\hat{\boldsymbol{B}}^\circ_{\boldsymbol{p}}$ in (3.10) yields the integrated GREG estimator proposed by Lemaître and Dufour (1987). Inserting $v_i = N_g^{-1}$, in turn, results in the integrated GREG estimator proposed by Nieuwenbroek (1993). This general definition considerably facilitates the comparison of these two integrated GREG estimators. In the following, we will no longer distinguish between a person- and household-level GREG estimator. Instead, we refer to an ordinary, in case of $v_i = 1$, or a generalized, in case of $v_i = N_g^{-1}$, integrated GREG estimator. We choose the terms *ordinary* and *generalized* to make references to underlying ordinary least squares (OLS) and generalized least squares (GLS) methods to determine the parameters in the integrated regression model.

3.1.5 Comparing an Ordinary and a Generalized Integrated GREG Estimator

Obviously, an ordinary and a generalized integrated GREG estimator differ with respect to the variance component in the assisting model ξ. In the ordinary case, with $v_i = 1$, the variance is assumed to be constant and thus homoscedastic. In contrast, the generalized case, with $v_i = N_g^{-1}$, entails that the variance of the variable of interest decreases with the household size and is thus heteroscedastic. The model properties of the estimated coefficients depend on the true underlying variance structure of the variable of interest. Provided that the true variance is heteroscedastic, a generalized coefficient is more efficient than an ordinary one, but no longer efficient in the sense of the Gauss-Markov theorem. It is only asymptotically efficient (cf. von Auer, 2007, p. 379).

In the literature, there is disagreement with respect to the efficiency of an ordinary ($v_i = 1$) and a generalized ($v_i = N_g^{-1}$) integrated GREG estimator. Wu et al. (1997) indicated that a generalized integrated coefficient minimizes the variance under cluster sampling. Their argument relies on the theory of optimal regression as introduced by Montanari (1987) (see Section 2.3.1). A similar explanation was given by Steel and Clark (2007, Theorem 1). However, both Wu et al. (1997) and Steel and Clark (2007) ignored the integrated property (see Section 3.1.1). Consequently, the design-based variance is minimized by a GREG estimator utilizing $\boldsymbol{x_g}$ instead of the integrated auxiliary vector $\boldsymbol{x}_{\boldsymbol{g}}^{\circ}$ as claimed by the mentioned authors. This result strongly depends on the variance formula of a person-level GREG estimator under cluster sampling, as we will discuss in detail in Chapter 6.

In contrast, Estevao and Särndal (2006) stated that a ordinary integrated GREG estimator (with $v_i = 1$) has a smaller variance, because the person-level residuals are based on a more proper regression and thus have a smaller magnitude than their ordinary counterparts. In a simulation study (Section 3.4), we verify which integrated GREG estimator is more efficient.

3.1.6 Integrated Weighting According to Estevao and Särndal (2006)

Estevao and Särndal (2006) introduced a different concept of integrated weighting in the context of two-stage cluster sampling. We consider their arguments with respect to single-stage cluster sampling with households sampled at the first stage and selecting all persons within a household at the second stage. Estevao and Särndal (2006) claimed that integrated weights are characterized by a convenient relationship. According to them, the person- and household-level weights do not necessarily have to be consistent, they only have to be related. It is assumed that the auxiliaries are available at both levels. Let $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{iQ})^T$ be the auxiliary vector for person i, as defined before. Let $\boldsymbol{a_g} = (a_{g1}, \ldots, a_{gK})^T$ be the auxiliary vector for household g. The known population totals are given by $\boldsymbol{T_x}$ and $\boldsymbol{T_a}$. Estevao and Särndal (2006) then defined the following combined person- and household-level vectors

$$\begin{aligned}\tilde{\boldsymbol{x}}_{\boldsymbol{i}} &= (\boldsymbol{x_i}^T, N_g^{-1}\boldsymbol{a_g}^T)^T \\ \tilde{\boldsymbol{a}}_{\boldsymbol{g}} &= (\sum_{i\in U_g} \boldsymbol{x_i}^T, \boldsymbol{a_g}^T)^T.\end{aligned} \tag{3.12}$$

Accordingly, the auxiliary vectors $\boldsymbol{x_i}$ and $\boldsymbol{a_g}$ are combined to vectors of dimension $(Q+K)$ by assigning the information from one level to the respective other level. Based on the constructed auxiliary vectors (3.12), Estevao and Särndal (2006) specified three integrated calibration estimators. The comparability of these calibration estimators with the integrated GREG estimators introduced in the previous Sections 3.1.2 and 3.1.3 is guaranteed, as calibration weights are asymptotically equal to GREG weights (see Section 2.3.4).

In the first proposed calibration estimator, the weights w_i^{ES1} are calibrated at the person level to satisfy the constraints

$$\sum_{i\in s_p} w_i^{\text{ES1}}\tilde{\boldsymbol{x}}_{\boldsymbol{i}} = \sum_{i\in U_p} w_i^{\text{ES1}}\begin{pmatrix}\boldsymbol{x_i}\\ N_g^{-1}\boldsymbol{a_g}\end{pmatrix} = \begin{pmatrix}\boldsymbol{T_x}\\ \boldsymbol{T_a}\end{pmatrix}.$$

Superscript ES1 refers to the first estimator proposed by Estevao and Särndal. The resulting calibration weights w_i^{ES1} can differ within a household, because of the original individual information received in the calibration. The household weights are then computed as the mean value of the person-level weights

$$w_g^{\text{ES1}} = \sum_{i\in U_g} N_g^{-1} w_i^{\text{ES1}}.$$

Because of differing person weights, the weights at the person and at the household level do not necessarily have to be equal. Hence, consistent estimates between both levels are not guaranteed. In addition, the household weights w_g^{ES1} no longer satisfy the person-level constraints: $\sum_{g\in U_h} w_g^{\text{ES1}}\boldsymbol{x_g} \neq \boldsymbol{T_x}$. Consistent results are only ensured if $\boldsymbol{x_i}$ in (3.12) is replaced by the constant household mean values $N_g^{-1}\boldsymbol{x_i}$. The weights w_i^{ES1} then simplify to the integrated weights suggested by Lemaître and Dufour (1987), defined in (3.3), with additional household-level auxiliaries.

In the second proposed estimator, the weights w_g^{ES2} are calibrated at the household level to satisfy the constraints

$$\sum_{g\in s_h} w_g^{\text{ES2}}\tilde{\boldsymbol{a}}_{\boldsymbol{g}} = \sum_{g\in U_h} w_g^{\text{ES2}}\begin{pmatrix}\sum_{i\in U_g}\boldsymbol{x_i}\\ \boldsymbol{a_g}\end{pmatrix} = \begin{pmatrix}\boldsymbol{T_x}\\ \boldsymbol{T_a}\end{pmatrix}.$$

Superscript ES2 refers to the second estimator proposed by Estevao and Särndal. The person weights are computed by

$$w_i^{\text{ES2}} = w_g^{\text{ES2}} \qquad \text{for all } i\in U_g.$$

Since the person weights within a household and the household weight itself are equal, consistency is ensured. Obviously, this proposed estimator is equivalent to the integrated GREG estimator suggested by Nieuwenbroek (1993) with additional person-level auxiliaries.

The third proposed estimator is computed in a two-step procedure. Superscript ES3 refers to the third estimator proposed by Estevao and Särndal. In the first step, the person weights w_i^{ES3} are calibrated to satisfy solely the person-level constraints

$$\sum_{i \in s_p} w_i^{\text{ES3}} \boldsymbol{x_i} = \boldsymbol{T_x}.$$

In the second step, the household-level weights w_g^{ES3} are computed to satisfy the constraints

$$\begin{pmatrix} \sum_{g \in s_h} w_g^{\text{ES3}} \boldsymbol{a_g} \\ \sum_{i \in U_g} w_i^{\text{ES3}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{T_a} \\ N_g w_g^{\text{ES3}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{T_x} \\ \boldsymbol{T_a} \end{pmatrix}.$$

This implies that every household g in the sample s_h imposes a unique constraint, which results in cumbersome computations in the case of household surveys.

With respect to the objective of this thesis, ensuring consistent estimates at the person and household level, we conclude the following:

- The first proposed estimator ensures consistent person- and household-level estimates only if the person-level auxiliaries are replaced by their household mean values. This results in the integrated GREG estimator introduced in Section 3.1.2.

- The second proposed estimator is equivalent to the integrated GREG estimator with households as basis introduced in Section 3.1.3.

- The third proposed estimator is not feasible in the context of household surveys.

Consequently, we do not pursue the approach of Estevao and Särndal (2006) hereinafter.

3.2 Consequences of Integrated Weighting

In this section, we adduce the consequences of the strict requirement of equal weights for all persons within a household and the household itself. We differentiate between the consequences of the replacement of the original auxiliaries with constructed household mean values (see Section 3.2.1) and the consequences due to the one-to-one weight assignment between the estimation levels (see Section 3.2.2). On the basis of these consequences, we deduce expectations about the performance of point and variance estimates of integrated weighting. The focus of this chapter lies on estimating person characteristics, as we expect the most consequences from the restriction of equal weights at this level.

To underpin our argumentation, we compare the integrated GREG estimator and naïve GREG estimator. The latter uses the original auxiliary information and does not ensure consistency. We use the synthetic data set AMELIA, which is based on EU-SILC (cf. Burgard et al., 2017). To

Table 3.1: Auxiliary variables I

Variable	Description
`sex`	Sex with two categories (male, female)
`age`	Age classes with four categories (younger than 19, 20-39, 40-59, 60 and older)
`ms`	Marital status with four categories (unmarried, married, separated or divorced, widowed)

reduce computational burden, we use the data of only one out of four regions. Then our population consists of approximately 2.6 million persons and 0.9 million households. The auxiliaries of the integrated and naïve GREG estimator are given in Table 3.1.

It is noteworthy that all numbers presented in this section refer to the population. A simulation study with random draws of samples to study the sampling distribution of the estimators follows in Section 3.4.

3.2.1 Consequences of the Replacement of Original Auxiliaries with Constructed Household Mean Values

In addition to the integrated GREG estimator requiring an additional auxiliary variable to ensure the integrated property, the difference between a naïve and an integrated GREG estimator is that the latter uses constructed household mean values instead of the original individual information. Obviously, the aggregation of the individual auxiliaries to a higher level causes some reduction of sample information. Sometimes, the technique of aggregation is used for reasons of disclosure control of sensitive data (cf. Bethlehem et al., 1990).

3.2.1.1 Increased Number of Outcome Values

Through the construction of household means, the outcome values of the original categories might be redistributed within a household. In explanation, suppose that a household g contains three household members: mother, father, and daughter. The original auxiliary vector for the variable `sex` is then given by $(x_1 = 0, x_2 = 1, x_3 = 0)^T$. In contrast, in the integrated weighting approach, the value $x_2 = 1$ will be redistributed to all other household members: $(\bar{x}_1 = 1/3, \bar{x}_2 = 1/3, \bar{x}_3 = 1/3)^T$. Whereas the original variable `sex` has only two outcome values, 0 or 1, the number of possible values in the integrated approach increases with the number of household members: for a single-person household it is 0 or 1; for a two-person household 0, 1/2, 1; for a three-person household 0, 1/3, 2/3, 1; for a four-person household 0, 1/4, 2/4, 3/4; and so on. Table 3.2 shows that the number of possible outcome values for the constructed auxiliaries $\bar{\boldsymbol{x}}_i$ significantly exceeds the number of possible outcome values for the

Table 3.2: Number of possible outcome values for x_i and $\bar{x}_i$

	age1	age2	age3	age4	sex0	sex1	ms1	ms2	ms3	ms4
x_i	2	2	2	2	2	2	2	2	2	2
$\bar{x}_i$	43	42	40	39	36	36	45	43	26	24

original auxiliary information $\boldsymbol{x}_i$. The increased number of outcome values might increase the range of the integrated weights and thus affect the efficiency. We evaluate this expectation, and all following expectations, in the simulation study in Section 3.4.

3.2.1.2 Ignoring the Heterogeneity within a Household

Because of the replacement of the original person-level auxiliaries with a constructed vector of household mean values, the variances of the auxiliaries are changed. The covariances with a variable of interest also differ. This becomes evident when decomposing the variance of an arbitrary original auxiliary vector $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ into the within and the between variance

$$\begin{aligned} V(\boldsymbol{x}) &= (N-1)^{-1} \sum_{i \in U_p} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2 \\ &= \underbrace{(N-1)^{-1} \sum_{i \in U_p} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})^2}_{\text{within variance}} + \underbrace{(N-1)^{-1} \sum_{i \in U_p} (\bar{\boldsymbol{x}}_{\boldsymbol{i}} - \bar{\boldsymbol{x}})^2}_{\text{between variance}}. \end{aligned}$$

Analogously, the covariance with an arbitrary variable of interest $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ can be decomposed into

$$\begin{aligned} \operatorname{Cov}(\boldsymbol{x}, \boldsymbol{y}) &= (N-1)^{-1} \sum_{i \in U_p} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})(y_i - \bar{y}) \\ &= \underbrace{(N-1)^{-1} \sum_{i \in U_p} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{i}})(y_i - \bar{y})}_{\text{within covariance}} + \underbrace{(N-1)^{-1} \sum_{i \in U_p} (\bar{\boldsymbol{x}}_{\boldsymbol{i}} - \bar{\boldsymbol{x}})(y_i - \bar{y})}_{\text{between covariance}}. \end{aligned}$$

The respective second term on the right-hand side describes the variance of an arbitrary integrated auxiliary vector $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_N)^T$

$$V(\bar{\boldsymbol{x}}) = (N-1)^{-1} \sum_{i \in U_p} (\bar{\boldsymbol{x}}_{\boldsymbol{i}} - \bar{\boldsymbol{x}})^2$$

and the integrated covariance

$$\operatorname{Cov}(\bar{\boldsymbol{x}}, \boldsymbol{y}) = (N-1)^{-1} \sum_{i \in U_p} (\bar{\boldsymbol{x}}_{\boldsymbol{i}} - \bar{\boldsymbol{x}})(y_i - \bar{y}).$$

As a consequence, the integrated approach captures only the between variance and between covariance. Ignoring the within variance and within covariance implies that the heterogeneity of the persons within a household is not taken into account. To assess the impact of ignoring the within variance, we compute the share of the within and between variance on the total variance. We continue with the same proceeding for the covariance with income (abbreviated with `inc`) as the variable of interest. Table 3.3 depicts that neither the within variance nor the within covariance (left columns) disregarded in integrated weighting are negligible. Additionally, the within variance and within covariance exceed the between variance and between covariance for any auxiliary variable. Hence, the integrated weighting approach does not exploit all available auxiliary information.

Table 3.3: Share of the within and between variance or covariance on the total variance or total covariance with income as variable of interest

	within	between	within	between
	variance		covariance	
age1	0.67	0.33	0.71	0.29
age2	0.63	0.37	0.68	0.32
age3	0.62	0.38	0.65	0.35
age4	0.52	0.48	0.56	0.44
sex0	0.74	0.26	0.75	0.25
sex1	0.74	0.26	0.75	0.25
ms1	0.64	0.36	0.69	0.31
ms2	0.62	0.38	0.68	0.32
ms3	0.64	0.36	0.71	0.29
ms4	0.65	0.35	0.75	0.25

Both the variance and covariance affect the stability of the estimated coefficients. The higher the variation of the auxiliaries, the more stable the projection onto the space spanned by the auxiliaries. Therefore, we expect that the integrated coefficients vary more compared to the coefficients resulting from a naïve GREG estimator. Moreover, the disregarded person-level variation might also affect the efficiency of the estimators, because even if the GREG estimator is model-assisted, and thus its design-based properties do not depend on the correctness of the model, its efficiency relies on the strength of the relationship between the variable of interest and the auxiliaries. These expectations will be validated in the simulation study.

3.2.2 Consequences of the One-to-One Weight Assignment

An interesting question is what consequences are caused by the one-to-one weight assignment from the person to the household level. One obvious disadvantage of the one-to-one weight assignment is that the same auxiliary information is demanded at the person and at the household level. Thus, the explanatory power of level-specific variables of interest is ignored.

Moreover, because of the one-to-one weight assignment, the integrated approach tacitly assumes that the strength of the relationship between the variable of interest and the auxiliary variables are identical at both levels. However, Robinson (1950) showed that the correlations for the same variables can be different at the individual level than at the aggregated level. This phenomenon is known as **ecological fallacy**. Misleading results are generated if the causes of variation between aggregated data differ from the causes of variation within the aggregated data (cf. Gelman et al., 2001, p. 110). Table 3.4 confirms that the correlations of `inc` with the original individual auxiliaries (left column) and with the household mean values (right column) differ. In particular, the correlations with the original auxiliaries considerably exceed the ones with the household mean values for every single variable. Therefore, we expect an efficiency loss for the integrated approach, because its coefficients are based on Cor$(\bar{\boldsymbol{x}}_i, \boldsymbol{y}_i)$ instead of on Cor$(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

Table 3.4: Correlations of the auxiliaries and `inc`

	Cor$(\boldsymbol{x}_i, \boldsymbol{y}_i)$	Cor$(\bar{\boldsymbol{x}}_i, \boldsymbol{y}_i)$
age1	-0.23	-0.12
age2	0.14	0.07
age3	0.15	0.09
age4	-0.09	-0.06
sex0	0.10	0.05
sex1	-0.10	-0.05
ms1	-0.16	-0.08
ms2	0.12	0.06
ms3	0.05	0.02
ms4	0.03	0.01

3.3 Empirical Evidence in the Literature

Before validating the previously discussed consequences of integrated weighting using a simulation study, we review the empirical evidence on integrated weighting found in the literature. The first four empirical studies are based on one single sample. The last two studies are based on repeatedly drawn samples from a fixed population.

Lemaître and Dufour (1987) compared the estimates of person characteristics obtained from an ordinary integrated GREG estimator with that from a classical post-stratification estimator using the 1981 Canadian LFS. For household characteristics, the reference post-stratification estimator used the principal person method presented in Section 3.1.2. The Canadian LFS is a monthly rotating panel survey of approximately 48,000 households with 10 geographic strata. Estimation was carried out for six different provinces of Canada with 24 sex by age groups as auxiliaries. The variables of interest at the person level consisted of the number of persons employed, unemployed, and unattached. The number of economic families, which comprise all persons in a household related by blood, marriage, or adoption, provided a household characteristic. The empirical study revealed that even if the integrated weights were more spread compared to those from the post-stratification estimator, no substantially different point estimates resulted. With respect to the precision of person characteristics estimates, almost no differences were realized. However, substantial efficiency gains were achieved for economic families. On the basis of these results, Lemaître and Dufour (1987) concluded that integrated weighting could be implemented with little or no loss of efficiency for estimates of person-level characteristics.

Using the Dutch Regional Income Survey (RIS), van den Brakel (2013, 2016) contrasted the efficiency of an ordinary integrated, a naïve GREG, and a Horvitz-Thompson estimator. The RIS is selected through stratified random sampling without replacement. Eighteen indicator variables of household type, different cross-classifications of age by gender, and age by gender by marital status served as auxiliaries. Variables of interest included the mean income of households and of persons as well as the income distribution of households in 10 classes where the categories are based on deciles. Point estimates and their corresponding standard errors were computed for three municipalities of different sample sizes (171,400, 46,300, and 2,500 persons) and for three subsequent years (2006, 2007, and 2008). When the efficiencies of Horvitz-Thompson and both GREG estimators are compared, the empirical results showed that the differences were small for municipalities with larger sample sizes. For smaller samples, the use of auxiliary information increased the precision of the GREG estimators. Comparing an integrated and a naïve GREG estimator, it becomes apparent that the latter is slightly less precise when estimating the total household income. The opposite is true when estimating person-level income. From that, van den Brakel (2013, 2016) concluded that the assumed variance structure of an ordinary integrated GREG estimator better fits to household- than to person-level characteristics.

It should be remarked that even if Lemaître and Dufour (1987) utilized a post-stratification estimator, their results are directly comparable to the results given by van den Brakel (2013, 2016) using a naïve GREG estimator. The reason for this is that only categorical auxiliaries are included. Therefore, post-stratification and the linear naïve GREG estimator are equivalent (cf. Zhang, 2000).

Isaki et al. (2004) used the 1990 U.S. Census of Population and Housing to adapt quadratic programming with the original weighting method implemented at that time. When estimating person characteristics, original raking was applied. When estimating household characteristics, the principal person method was used. As mentioned in Section 3.1.3, quadratic programming is equivalent to a calibration estimator with a truncated distance function. Thus, except for the bounds, it is comparable with a generalized GREG estimator. Both methods under consideration were assessed in terms of the agreement between estimates and census counts. The estimates were produced based on data out of one weighting area of the census containing 8,034 households and 25,145 persons. The variables of interest consisted of several categorical person and household characteristics. The empirical results revealed that for household characteristics, both methods perform similar. For person-level characteristics, the quadratic programming estimates better fit the census counts than the raking estimates. In their conclusion, Isaki et al. (2004) emphasized the computational feasibility of producing one integrated weight rather than separate person and household weights.

Branson and Wittenberg (2014) presented estimates from their proposed cross-entropy estimation approach using the South African Household Survey, the LFS, and the General Household Survey for the years from 1994 up to 2007. The post-stratification estimator that was currently implemented, served as a benchmark estimator. All household surveys at hand were sampled by means of a two-stage cluster design, with regional areas selected via simple random sampling[1] at the first stage and households selected via stratification at the second stage. Neither the number of sampling units nor the auxiliary information was mentioned by the authors. However, it can be presumed that the same auxiliaries were applied as available for the original weighting procedure: provinces, age, sex, race, and urban or rural. The variables of interest comprised the number of persons and households, age, sex, race, and regional variables. From the results, Branson and Wittenberg (2014) claimed their cross-entropy estimates ensure consistent results and form a smooth time series, whereas the original estimates exhibit jumps.

Using a simulation study as their basis, Steel and Clark (2007) investigated the relative root MSE (see Section 2.2.2 for a definition) of a naïve, an ordinary, and a generalized integrated GREG estimator. The 2001 Australian Population Census consisting of 187,178 households and the 1995 Australian National Health Survey consisting of 210,132 persons served as the population. From these populations, 5,000 samples of different sizes were drawn via simple single-stage cluster sampling of households. The auxiliaries consisted of 12 indicator variables of sex-by-age classes. The variables of interest included six variables out of the census (employed, employed female, income, low income, hours worked) and five health variables. For the health variables, both integrated GREG estimators performed slightly better than the naïve GREG estimator for a sample size of 1,000 households. The reverse is true for the census variables. The precision improvement of both integrated GREG estimators with respect to census variables increased with the sample sizes. The efficiency loss with respect to health variables decreased when increasing the sample size. For regional estimates, both integrated GREG estimators were slightly worse in all cases. Steel and Clark (2007) claimed that there is little or no loss associated with the practical benefit of integrated weighting. This conclusion is in accordance with the conclusion drawn by Lemaître and Dufour (1987).

Wu et al. (1997) contrasted an ordinary and a generalized integrated GREG estimator. A simulation study was implemented with data from the 1990 Canadian LFS of Newfoundland consisting of 9,152 persons. A two-stage cluster design was realized with selecting PSU via probability proportional to size at the first stage and with selecting dwellings via simple random sampling at the second stage. All persons within a dwelling were included into the sample. One thousand

[1] In South Africa, a sampling frame containing all households is available.

samples of approximately 1,000 persons were drawn from the population. Twenty-four different sex-by-age groups served as auxiliaries. The variables of interest included the number of persons employed and unemployed in single-person households and the number employed in households of four and more persons. For both single-person variables, the ordinary integrated GREG estimator performs slightly worse than the generalized one. No difference occurred for the number of employed for households of size four and more. These results confirmed their theoretical expectation based on arguments of the theory of optimal estimators that the generalized GREG estimator was preferable in terms of asymptotic efficiency. Both integrated estimators had a negligible bias.

Comparing the presented studies, it becomes apparent that the point estimates from a naïve and an integrated GREG estimator were only contrasted by Lemaître and Dufour (1987). They found no considerable differences. With respect to the efficiency, Lemaître and Dufour (1987) and van den Brakel (2013, 2016) agreed that the integrated GREG estimator is more efficient when estimating household characteristics. However, the universal recommendation of an integrated GREG seems surprising, as van den Brakel (2013, 2016) and Steel and Clark (2007) found less efficient results with respect to income, health variables, and regional estimates.

A key limitation of all the mentioned studies is that none stated the additional auxiliary variable required to ensure the integrated property introduced in Section 3.1.1. The number of variables of interest is very limited in all studies except for Steel and Clark (2007) and Isaki et al. (2004). Also, the types of variables of interest are limited. Almost all variables are dichotomous except for income. Moreover, the population from which Wu et al. (1997) drew the samples is very small. Wu et al. (1997) and Lemaître and Dufour (1987) used the Canadian LFS as data base. However, although the latter computed one single estimate for one single sample, the former ran a simulation study with 5,000 samples. Both studies included the same auxiliary variables. Nevertheless, their results are not directly comparable, as the focus of the studies differs: a comparison of an integrated and a post-stratification estimator versus a comparison of an ordinary and a generalized integrated GREG estimator.

3.4 Simulation Study

The following Monte-Carlo (MC) simulation study compares the performance of the integrated GREG estimators and a naïve GREG estimator in order to evaluate the aforementioned consequences of integrated weighting. See Table 3.5 for the estimators under consideration.

3.4.1 Simulation Setup

The simulation study is based on the aforementioned synthetic population AMELIA (cf. Burgard et al., 2017). The auxiliaries consist of an intercept and 18 indicator variables (see Table 3.6). Within the integrated GREG estimators, we also include the additional auxiliary, N_g^{-1},

Table 3.5: Estimators under consideration

Estimators	Description
GREG	Naïve GREG estimator determined by Definition 2
GREG2	Naïve GREG estimator determined by Definition 2 with household size as additional auxiliary
INT1	Ordinary integrated GREG estimator determined by Definition 3 with $v_i = 1$
INT2	Generalized integrated GREG estimator determined by Definition 3 with $v_i = N_g^{-1}$
INT1b	Ordinary integrated GREG estimator determined by Definition 3 with $v_i = 1$ without the integrated variable N_g^{-1}
INT2b	Generalized integrated GREG estimator determined by Definition 3 with $v_i = N_g^{-1}$ without the integrated variable N_g^{-1}

required to ensure the integrated property. The choice of solely categorical variables induces that the GREG estimator is equivalent to a post-stratification estimator (cf. Zhang, 2000).

Table 3.6: Auxiliary variables II

Variable	Description
intercept	Intercept
sex	Sex with two categories (male, female)
age	Age classes with four categories (younger than 19, 20-39, 40-59, 60 and older)
ms	Marital status with four categories (unmarried, married, separated or divorced, widowed)
sex_age	Cross-classifications of age by sex with four categories
ms_sex	Cross-classifications of marital status by sex with four categories

We chose three different types of variables of interest (see Table 3.7). The cross-classifications with the household size are included because, as mentioned in Section 3.3, Lemaître and Dufour (1987) and van den Brakel (2013, 2016) emphasized the superiority of integrated weighting with respect to household-level characteristics. We deliberately chose variables on different scales: metric, dichotomous, and categorical. Income, moreover, is a skewed variable. The broad range of different scales and different degrees of skewness induces varying difficulties to produce unbiased and efficient estimates. Therefore, a comprehensive comparison of the performance of an integrated and a naïve GREG estimator is guaranteed.

Table 3.7: Variables of interest at the person level

Variable	Description
Classical person-level characteristics	
inc	Personal income
soc	Social income
sel	Self-employment dummy
act	Activity status with three categories (at work, unemployed, inactive persons)
Cross-classification with the household size	
inc_hs	Cross-classification of personal income by household size with six categories
Cross-classification with the auxiliaries	
bene_age	Cross-classification of unemployment benefits by age with four categories

We drew $R = 1000$ MC samples via simple random sampling. We chose a larger sample size of $m = 1500$ households and a smaller sample size of $m = 200$. The average MC sample size of persons is $\sum_{r=1}^{1000} n_r = 4333$ in case of $m = 1500$ sampled households and $\sum_{r=1}^{1000} n_r = 577$ for $m = 200$ sampled households.

Two aspects are relevant when evaluating the different methods: point and variance estimates. First, we introduce some quality measures for point estimates based on the quality criteria presented in Section 2.2.2. Suppose $\hat{T}_r$ as the resulting total estimate for the r-th MC replication with $r = 1, \ldots, R$. Define $E^{\text{MC}}(\hat{T}) = R^{-1} \sum_{r=1}^{R} \hat{T}_r$, where the quantity $E^{\text{MC}}(\hat{T})$ denotes the empirical expectation of the estimator $\hat{T}$. Let T be the true value. Then, the empirical relative bias (RB) of $\hat{T}$ is given by

$$\text{RB}(\hat{T}) = \frac{E^{\text{MC}}(\hat{T}) - T}{T}.$$

The RB measures the mean difference of the estimator from the true value in relation to the true value itself. The empirical mean squared error (MSE) is defined as

$$\text{MSE}(\hat{T}) = \left(E^{\text{MC}}(\hat{T}) - T\right)^2.$$

The empirical relative root mean squared error (RRMSE) is given by

$$\text{RRMSE}(\hat{T}) = \sqrt{\frac{\left(E^{\text{MC}}(\hat{T}) - T\right)^2}{T^2}}.$$

The MSE and RRMSE take into account both the bias and the variability of the estimates. The advantage of the RRMSE compared to MSE is that it allows a relative comparison between different variables. However, an issue arises is if the denominator in RRMSE is close to zero. Because of this issue and because we are not interested in a comparison between the variables, but between the estimators, we prefer the MSE as the quality measure in the following.

We continue by introducing quality measures for the variance estimates. Suppose $\hat{V}_r(\hat{T})$ is the variance estimate for the r-th MC replication with $r = 1, \dots, R$. Define $E^{\text{MC}}[\hat{V}(\hat{T})] = R^{-1} \sum_{r=1}^{R} \hat{V}(\hat{T})_r$ where the quantity $E^{\text{MC}}[\hat{V}(\hat{T})]$ denotes the empirical expectation of the variance estimator $\hat{V}(\hat{T})$. Consider $V^{\text{MC}}(\hat{T}) = R^{-1} \sum_{r=1}^{R} [\hat{T}_r - E^{\text{MC}}(\hat{T})]^2$ as the empirical variance of $\hat{T}$. Then, the RB of the variance estimates is given by

$$\text{RB}[\hat{V}(\hat{T})] = \frac{E^{\text{MC}}[\hat{V}(\hat{T})] - V^{\text{MC}}(\hat{T})}{V^{\text{MC}}(\hat{T})}.$$

The RB, MSE and RRMSE produce one number for all MC replicates. Hence, we define a replication-specific quality measure for the variance estimates. The replication-specific relative bias of $\hat{V}(\hat{T})$ is determined by

$$\text{rsRB}_r[\hat{V}(\hat{T})] = \frac{\hat{V}_r(\hat{T}) - V^{\text{MC}}(\hat{T})}{V^{\text{MC}}(\hat{T})}.$$

The rsRB$_r$ measures the relative deviation of each r variance estimate from the empirical variance of the point estimator. We use $R = 10,000$ MC replicates to compute $V^{\text{MC}}(\hat{T})$, because the empirical variance of the point estimates $V^{\text{MC}}(\hat{T})$ converges slower than the empirical expectation of the variance estimates $E^{\text{MC}}[\hat{V}(\hat{T})]$.

3.4.2 Results on Weights and Regression Coefficients

We start by analyzing the weight distributions generated by the methods under consideration in order to evaluate whether the increased number of outcome values, as discussed in Section 3.2.1, might induce a wider range of the integrated weights. Figure 3.1 presents the weights at the person level divided by the design weights for all $R = 1000$ MC replications. Per method approximately 4.3 million data points are plotted. It becomes apparent that both integrated weights have a wider range and a higher variation compared to the weights of a naïve GREG estimator. A greater dispersion of the integrated weights was also found by Lemaître and Dufour (1987) and Rottach and Hall (2005). Actually, negative weights occur for both integrated GREG estimators. The GREG weights are positive throughout, even for the smaller sample size. Consequently, the weight distribution of an integrated GREG estimator is significantly influenced by the redistribution of the original auxiliaries to all household members.

Interestingly, the variation of the integrated person weights depends on the household size (see Figure 3.2). All households of size > 6 were collapsed to one category. Although the weights vary more for smaller households for INT1, the opposite is true for INT2. The reason is the reverse assumed variance structure of the variable of interest in the assisting model.

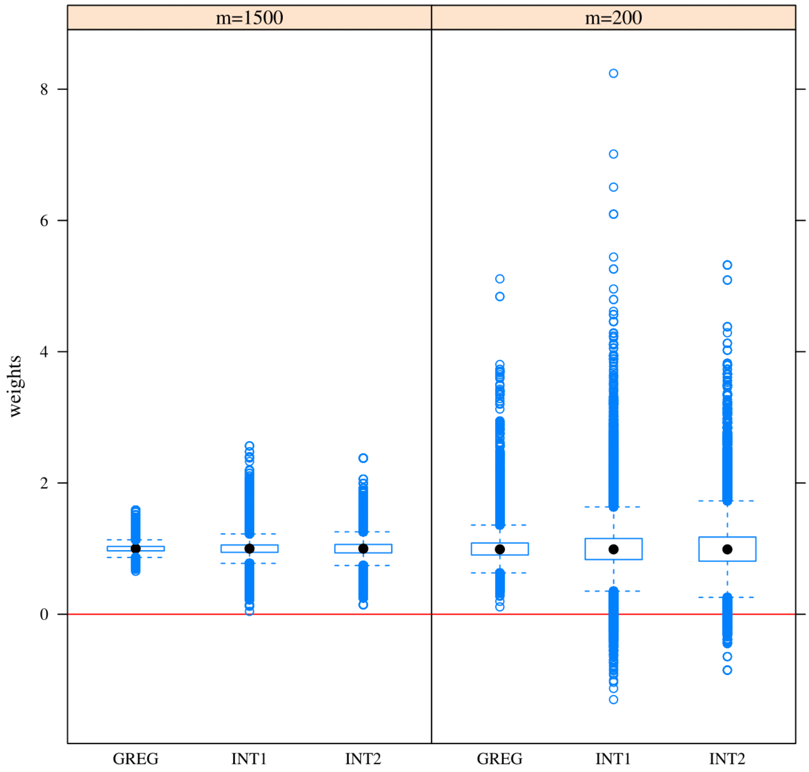

Figure 3.1: Boxplots for person weights

To evaluate whether the ignorance of the heterogeneity of the persons within a household (see Section 3.2.1) affects the regression coefficients, we continue by analyzing the regression coefficients of the estimators under consideration. Figure 3.3 opposes $\hat{\boldsymbol{B}}_r^{\text{GREG}}$, $\hat{\boldsymbol{B}}_r^{\text{INT1}}$ and $\hat{\boldsymbol{B}}_r^{\text{INT2}}$ for $= 1, \ldots, 1000$ MC replications. To center the coefficients around zero, we subtract the true regression coefficients. For all auxiliaries, the coefficients of INT1 and INT2 considerably vary more than the coefficients of GREG. It should be noted that a naïve GREG estimator does not require the additional variable N_g^{-1}; thus there are no boxes for it. This result confirms our expectation from Section 3.2.1.2 that the neglected within variance in the integrated weighting approach results in less-stable coefficients.

3.4.3 Results on Point Estimates

Table A.1 in Appendix A validates the property that all estimators under consideration are unbiased or nearly so. Table 3.8 summarizes the relative efficiency of the MSE for different variables of interest and for different sample sizes. For $m = 1500$, GREG performs similarly to INT1 and INT2 for person characteristics and slightly worse for variables related to household size. This result is in accordance with the empirical studies given in Lemaître and Dufour (1987) and van den Brakel (2013, 2016). For the smaller sample size $m = 200$, the efficiency

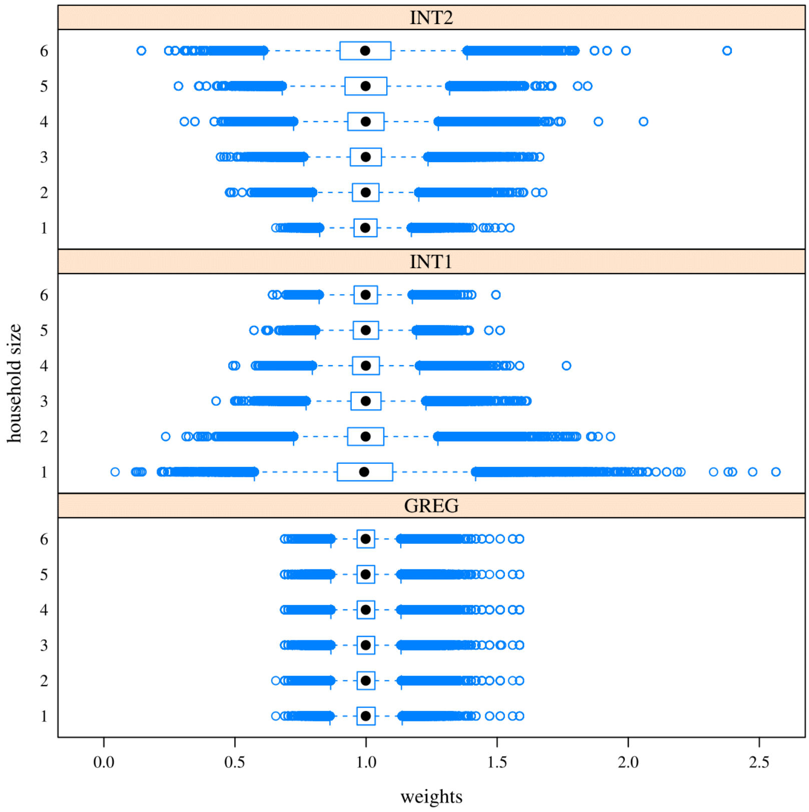

Figure 3.2: Boxplots for person weights by household size for $m = 1500$

gains of GREG relative to INT1 and INT2 increase. As INT1 and INT2 contain the integrated variable as an additional auxiliary, we include the household size, N_g, into the naïve GREG estimator for a fair comparison. We denote this estimator as GREG2. GREG2 outperforms INT1 and INT2 for all variables of interest including the variables related to household size. This relative improvement ranges from 7% to 49%. Therefore, when including the household size as an auxiliary into the naïve GREG estimator, the superiority of the integrated GREG estimator with respect to household-level characteristics as claimed by Lemaître and Dufour (1987) and van den Brakel (2013, 2016) vanishes. Our results indicate that the statement given by Lemaître and Dufour (1987), Steel and Clark (2007) and van den Brakel (2013, 2016) that the inefficiencies of integrated weighting would be limited, is valid only for larger sample sizes. It is important to note that $m = 200$ refers to the number of households. The average number of persons, on which the estimators are based, is $\bar{n} = 577$.

Table 3.9 contrasts the efficiency of INT1 relative to INT2. For $m = 1500$, INT2 performs better for almost all variables related to household size. This is in line with the findings of Wu et al. (1997) as discussed in Section 3.1.5. For person characteristics, there is no clear tendency for the superiority of one estimator. However, for the smaller sample size, INT1 seems to be slightly more efficient for most variables.

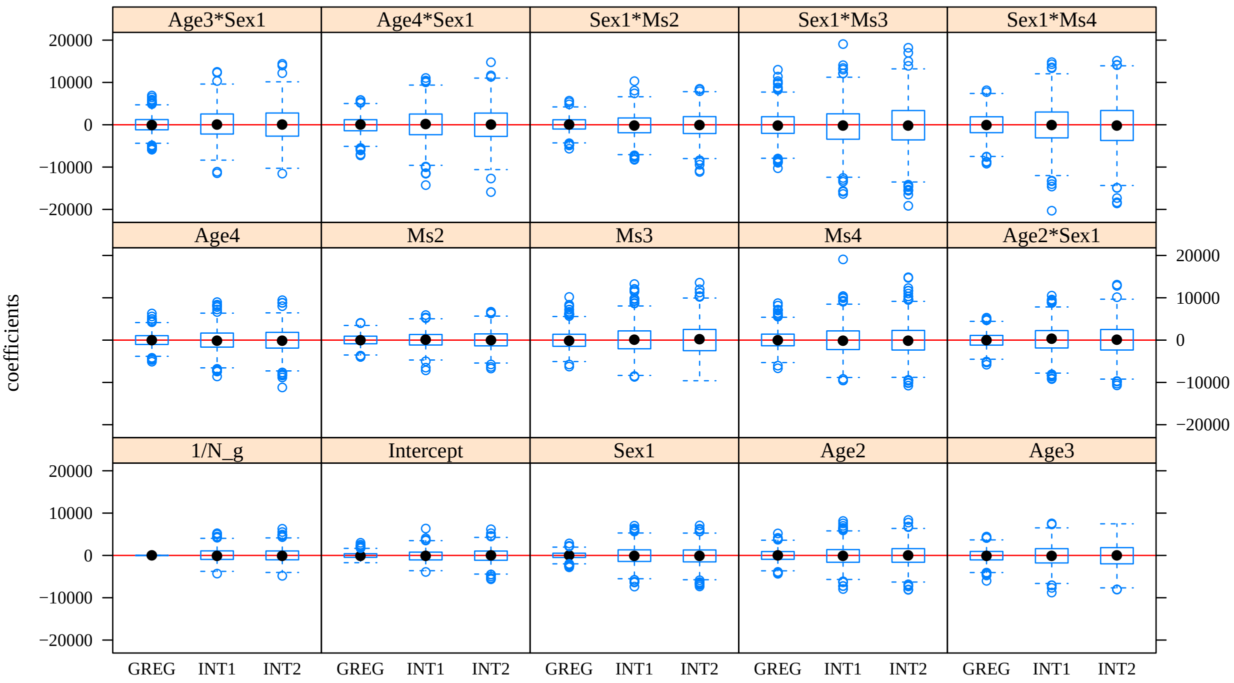

Figure 3.3: Boxplots for the coefficients $\hat{\boldsymbol{B}}_r^{\text{GREG}}$, $\hat{\boldsymbol{B}}_r^{\text{INT1}}$ and $\hat{\boldsymbol{B}}_r^{\text{INT2}}$ for $r = 1, \ldots, 1000$ and $m = 1500$

3.4.4 Results on Variance Estimates

None of the empirical studies presented in Section 3.3 examined the variance estimates of an integrated GREG estimator. Only the biases and efficiency of the point estimates were analyzed. The rsRB$_r$ for $r = 1, \ldots, 1000$ of the variance estimates are illustrated in Figure 3.4 The RB is indicated in green. For the smaller sample size (upper plots) and more skewed variables (bene and inc), the variance estimates are less precise than for the larger sample size (lower plots) and less skewed variables (act). This observation is valid for all estimators. The higher ranges of the RB and rsRB$_r$ for inc by hs and bene by age is caused by the decreased sample sizes in the cross-classifications.

For $m = 1500$, the performances of the variance estimators under consideration are quite similar. Except for inc_hs1, inc_hs2 and inc_hs6, GREG underestimates the empirical expectation of the variances, but produces fewer outliers. For $m = 200$, however, GREG achieves more precise variance estimates for most variables. Therefore, for smaller sample sizes the integrated GREG estimators produces less precise variance estimates compared to the naïve GREG estimator. This is in line with the results on the point estimates.

While exploring the consequences of integrated weighting, we generate several further simulation results. We especially aimed at detecting whether the replacement of the original auxiliaries might introduce some bias, which was not the case. The further simulation results can be found

Table 3.8: Relative efficiency of the MSE of point estimates

	m=1500				m=200			
	$\frac{\text{INT1}}{\text{GREG}}$	$\frac{\text{INT2}}{\text{GREG}}$	$\frac{\text{INT1}}{\text{GREG2}}$	$\frac{\text{INT2}}{\text{GREG2}}$	$\frac{\text{INT1}}{\text{GREG}}$	$\frac{\text{INT2}}{\text{GREG}}$	$\frac{\text{INT1}}{\text{GREG2}}$	$\frac{\text{INT2}}{\text{GREG2}}$
inc	1.00	1.00	1.00	1.00	1.04	1.04	1.02	1.02
soc	1.01	1.01	1.01	1.01	1.02	1.03	1.01	1.02
sel	1.00	1.00	1.00	1.00	1.01	1.02	1.00	1.01
act1	1.00	1.00	1.00	1.00	1.02	1.02	1.01	1.01
act2	1.00	1.00	1.00	1.00	1.02	1.03	1.01	1.02
act3	1.00	1.00	1.00	1.00	1.02	1.02	1.01	1.01
inc_hs1	0.94	0.90	1.13	1.08	1.03	0.90	1.24	1.10
inc_hs2	0.95	0.93	1.30	1.28	0.99	0.94	1.33	1.26
inc_hs3	1.00	1.00	1.38	1.36	1.03	1.02	1.45	1.44
inc_hs4	1.01	1.01	1.49	1.49	1.01	1.04	1.45	1.48
inc_hs5	0.98	0.99	1.10	1.11	1.00	1.04	1.07	1.11
inc_hs6	0.94	0.90	1.11	1.07	0.95	0.97	1.12	1.15
bene_age1	1.00	1.01	1.00	1.01	1.02	1.01	1.01	1.00
bene_age2	1.00	1.00	1.00	1.00	1.03	1.04	1.01	1.02
bene_age3	1.00	1.00	1.00	1.00	1.01	1.02	0.99	1.00
bene_age4	1.00	1.00	1.00	1.00	1.02	1.03	1.01	1.02

in Section A.2 in Appendix A. We addressed the topics of the estimation of subgroups and domains, the influence of equal weights on regressions and the influence of ecological fallacy in the integrated weighting approach.

Table 3.9: Relative efficiency of the MSE of point estimates of INT1 and INT2

	m=1500	m=200
	$\frac{\text{INT1}}{\text{INT2}}$	$\frac{\text{INT1}}{\text{INT2}}$
`inc`	1.00	1.00
`soc`	1.00	0.99
`sel`	1.00	0.99
`act1`	1.00	1.00
`act2`	1.00	0.99
`act3`	1.00	1.00
`inc_hs1`	1.05	1.13
`inc_hs2`	1.02	1.06
`inc_hs3`	1.01	1.01
`inc_hs4`	1.00	0.98
`inc_hs5`	0.99	0.96
`inc_hs6`	1.04	0.98
`bene_age1`	0.99	1.01
`bene_age2`	1.00	0.99
`bene_age3`	1.00	0.99
`bene_age4`	1.00	0.99

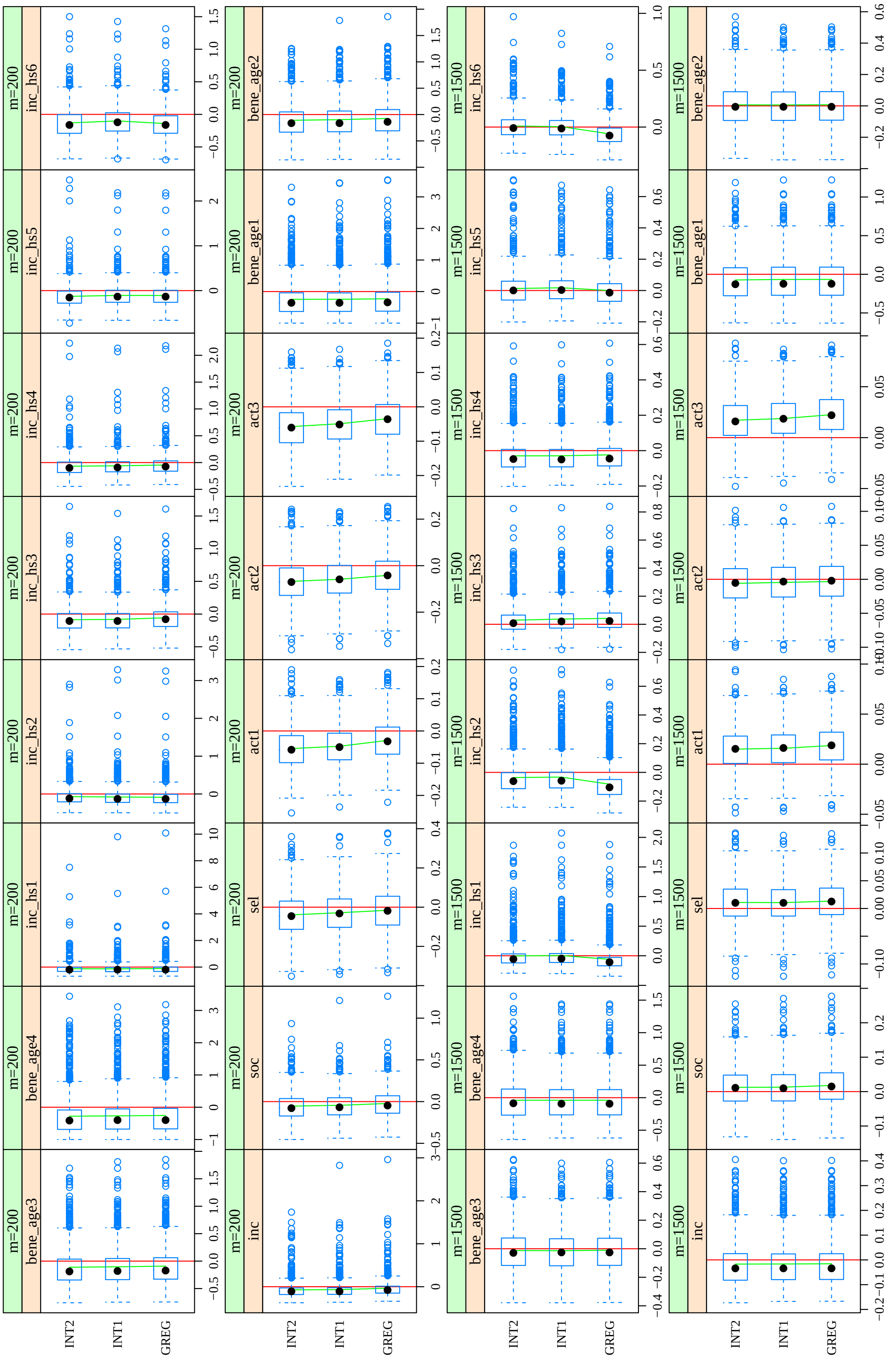

Figure 3.4: Relative Bias and replicate specific-relative bias of the estimated variances

3.5 Conclusion

This chapter introduced integrated weighting as the current method used by statistical offices to ensure consistent estimates at the person and household level. We emphasized an important property of integrated weighting that is essential to ensuring that the person weights sum up to the number of persons in the population and the household weights simultaneously sum up to the number of households in the population. This property, which we denote as the integrated property, is relevant only for integrated weighting, because the weights are calculated at one level and then assigned one-to-one to the other level. To the best of our knowledge, the integrated property is neglected in the literature.

To permit a comparison, we generalized the integrated GREG estimators introduced by Lemaître and Dufour (1987) and Nieuwenbroek (1993) into one single integrated GREG estimator derived at the person level with differing variance components (see Definition 3). To produce consistent estimates, the original auxiliary information is replaced by constructed household mean values. The person weights, which are equal for all household members, are then assigned one-to-one to the corresponding households. As a consequence of this approach, the number of outcome values of the auxiliaries increases, the heterogeneity of persons within a household is neglected by disregarding the within variance, and ecological fallacy can arise. The simulation study confirms that these consequences result in more spread weights, more varying coefficients, and a lower degree of explanation. Moreover, for smaller sample sizes, the point estimates are less efficient compared to a naïve GREG estimator. For larger sample sizes, our results are in accordance with the results given by Lemaître and Dufour (1987) and van den Brakel (2013, 2016). However, in contrast to the aforementioned authors, we include the integrated variable, which is essential for ensuring the integrated property. Therefore, the claimed superiority with respect to the estimation of household characteristics mentioned by Lemaître and Dufour (1987), van den Brakel (2013), van den Brakel (2016), and Steel and Clark (2007) vanishes when including the household size as auxiliary variable in the naïve GREG estimator. Indeed, when the household size is included, the naïve GREG estimator considerably outperforms the integrated GREG estimators. Furthermore, the variance estimator of an integrated GREG estimator is less efficient than that of a naïve GREG estimator for smaller sample sizes. To conclude, our results suggest that the consequences of integrated weighting outbalance that a naïve GREG estimator misses the fact that all household members are sampled as a cluster, as criticized for example by Steel and Clark (2007), Lemaître and Dufour (1987), and Wu et al. (1997, p. 101).

Therefore, to avoid the reported undesirable consequences of integrated weighting, we introduce in the following chapter two alternative weighting approaches that ensure consistent person- and household-level estimates without the strict requirement of equal weight of all household members and the household itself. The resulting individual person weights retain the individual patterns within a household and thus capture the heterogeneity in a household.

4 Alternative Weighting Approaches

In statistical offices, it is common practice to use integrated weighting, which enforces equal weights for all persons within the same household, to ensure consistent estimates. However, as demonstrated in detail in Section 3.2, the strict requirement of equal weights entails some undesirable consequences such as an increased number of outcome values of the auxiliary variables, the disregarded heterogeneities within a household, and possible problems induced by ecological fallacy. Therefore, in this chapter we introduce two weighting approaches as alternatives to integrated weighting, which are capable of both ensuring consistent person and household estimates and allowing for different weights of persons within a household.

The idea underlying our alternative weighting approaches is to constrain the consistency requirements to variables that are common to the person and household data set. By incorporating the common variables as additional auxiliaries, our alternative weighting approaches produce consistent estimates of these variables. Thus, consistency is ensured more directly and only for the relevant variables, instead of indirectly by aggregating the individual information per household. To implement such alternative weighting approaches, we adapt the method given by Renssen and Nieuwenbroek (1997), known from the literature on combining information from multiple independent surveys. However, there are major differences between our proposed alternative weighting approaches and the approach of Renssen and Nieuwenbroek (1997).

The main advantage of our alternative weighting approaches compared to integrated weighting is that the original individual rather than the constructed aggregated auxiliaries are utilized. Therefore, differing weights for the persons within a household are allowed that retain the individual pattern of the persons. Furthermore, our alternative weighting approaches consist of separate person- and household-level estimators, providing two further advantages. Firstly, the different calculation levels of person and household characteristics, which prevent the problems caused by ecological fallacy, are considered. Secondly, the variable selection process is more flexible because different auxiliary variables can be incorporated in the person-level estimator than in the household-level estimator. Finally, no additional auxiliary variable is required to enforce the integrated property. As a result, our alternative weighting approaches contradict the widespread opinion in the literature that equal weights are indispensable to ensure consistent estimates in household surveys.

Therefore, the aim of this chapter is to derive point and variance estimators of the proposed alternative weighting approaches. The chapter is organized as follows: Section 4.1 briefly reviews the literature on methods to combine information collected from independent multiple surveys. In Section 4.2, we introduce our two alternative weighting approaches to ensure consistency between person- and household-level estimates without the strict requirement of equal weights for all persons within the same household. To assess the impact induced by the consistency requirements, we decompose our proposed estimators into a naïve GREG estimator and an adjustment term capturing the effect of the consistency requirements. Section 4.3 discusses the GLS adjustment algorithm suggested by Zieschang (1986, 1990) and Merkouris (2004) which can also be adapted to guarantee consistent person- and household-level estimates. The GLS estimator serve as a benchmark for our proposed weighting approaches. For a better comparability with our proposed weighting approaches, we rewrite the GLS estimator as GREG estimator. Subsequently, we decompose the rewritten estimator into a naïve GREG estimator and an adjustment term capturing the consistency requirements. Section 4.4 emphasizes the difference between our proposed alternative weighting approaches and the GLS estimator. A simulation study in Section 4.5 contrasts the estimation performance of our proposed alternative weighting approaches, the GLS estimator, and integrated weighting. Section 4.6 contains concluding remarks.

For a better overview, Table 4.1 summarizes all discussed estimators presented in the following. At the end of Section 4.4, we complete this table with the corresponding formulas.

Table 4.1: Summary of the proposed and benchmark estimators I

Estimator	Abbreviation	Section
Proposed estimators		
First weighting approach	WA1	Section 4.2.1
Second weighting approach	WA2	Section 4.2.2
Benchmark estimators		
GLS estimator by Zieschang (1986, 1990)	ZIE	Section 4.3.1
GLS estimator by Merkouris (2004)	MER	Section 4.3.2

4.1 Methods to Combine Information from Independent Multiple Surveys

Since the idea of our alternative weighting approaches is based on methods to combine information from multiple independent surveys, we briefly review the literature on these methods. Zieschang (1986, 1990) explored a GLS adjustment algorithm to link the estimates of two independent surveys. The application focused on the Diary Survey and the Interview Survey as two separate components of the U.S. Consumer Expenditure Survey. The collected information obtained from each survey overlaps for some items such as tenure status or region of residence. To link the estimates of the overlapping variables, the auxiliary information of both surveys is pooled and additional linear constraints are imposed into the GLS estimator. Merkouris (2004) modified the GLS adjustment algorithm to account for different samples sizes. We return to the GLS adjustment algorithm in Section 4.3 since it can be adopted to ensure consistency between person- and household-level estimates.

To align the estimates for variables common to two independent surveys, Renssen and Nieuwenbroek (1997) introduced an adjusted GREG estimator. Common variables were defined as variables observed in two surveys for which the corresponding population totals are unknown. As estimators for the unknown common variable totals, Renssen and Nieuwenbroek (1997) suggested a weighted average of the estimates obtained from each of the independent surveys. By including the common variables as additional auxiliaries, the adjusted GREG estimator aligns the estimates of the respective common variables. We revive to the adjusted GREG estimator in Section 4.2 because our proposed alternative weighting approaches are based on the idea of including common variables as additional auxiliaries to ensure consistency.

Wu (2004) applied a pseudo-empirical likelihood (EL) approach to combine information from multiple independent surveys. The pseudo-EL approach is a nonparametric method, where estimation is based on a likelihood function (cf. Chen and Sitter, 1999). Wu (2004) showed that under suitable regularity conditions the EL approach is asymptotically equivalent to the GREG estimator, but with the advantage of producing only positive weights. The EL approach was also used by Kabzinska and Berger (2015) to align estimates from different surveys. Their proposed method differs from the EL approach suggested by Wu (2004) in that the inclusion probabilities are imposed into the linear constraints rather than into the likelihood function. However, we do not follow the EL approach in this thesis, because it does not provide an analytical expression for the weights and because it is asymptotically equivalent to the GREG estimator.

To enforce consistency among contingency tables of surveys estimates, Boonstra et al. (2003) and Houbiers (2004) introduced repeated weighting. Repeated weighting is a two-step procedure: In a first step, the contingency tables are estimated by means of a GREG estimator. In a second step, the table estimates are consecutively adjusted such that numerical consistency between the estimates is obtained (cf. Knottnerus and van Duin, 2006). Nevertheless, we do not pursue repeated weighting because the proposed method is applicable only for contingency tables. The main concern of this chapter, in turn, is to produce global weights that are suitable for all survey variables.

A considerable body of literature exists on the topic of combining information from multiple surveys accompanied with keywords such as multipurpose surveys, split questionnaires, or double sampling. However, this thesis focuses less on combining information from multiple surveys and more on ensuring consistency between person- and household-level estimates. Hence, we consider only the methods introduced by Zieschang (1986, 1990), Merkouris (2004), and Renssen and Nieuwenbroek (1997) in the following.

4.2 Alternative Weighting Approaches to Ensure Consistent Estimates

The objective of our proposed alternative weighting approaches is to ensure consistency between person- and household-level estimates without the strict requirement of equal weights for all persons within the same household as enforced by integrated weighting. Therefore, the individual patterns within a household are preserved. The underlying idea is to reduce the consistency requirements to variables that are common to the person and household data set. To implement such a weighting approach, we adopt the idea of Renssen and Nieuwenbroek (1997) to incorporate common variables into the estimation process. However, there are major differences between our proposed weighting approaches and the original method by Renssen and Nieuwenbroek (1997):

1) Renssen and Nieuwenbroek (1997) (and in general the literature on combining information from multiple surveys) considered **independent surveys** selected in separated sampling processes. In contrast, we deal with household surveys consisting of two highly dependent data sets, namely the person data set and the household data set, collected within one single sampling process.

2) A further considerable difference concerns the **definition of common variables**. Renssen and Nieuwenbroek (1997, p. 368) defined common variables as variables observed in two independent surveys for which the population totals are unknown. However, in the context of household surveys, it is hardly conceivable that the same item is simultaneously requested in the person and in the household questionnaire. Instead, it is more prevalent that for some person characteristics, their corresponding values at the household level are of interest. For such person characteristics, the per-household aggregated variables are computed and supplementarily added to the household data set. Plausible examples include household income, purchases, or the number of employees in a household. Therefore, we define common variables as variables available at different aggregation levels, once in their initial form at the person level and once in its aggregated form at the household level. It is notable that in our context practical complications induced by discrepancies between varying definitions, sampling modes and reference periods as arising in the context of multiple surveys, are irrelevant.

3) Lastly, Renssen and Nieuwenbroek (1997) assumed that the independent surveys at hand refer to the same **target population**. In contrast, household surveys provide information on both the population of persons and the population of households. As a consequence, in our context consistency requirements target the estimation at different aggregation levels.

Considering these differences, we propose two alternative weighting approaches that ensure consistency between the person- and household-level estimates. Because of these differences, the proposed point and variance estimators differ from the estimators by Renssen and Nieuwenbroek (1997). The differences in point and variance estimators are discussed in Section 4.4.

The remainder of this section is organized as follows: Section 4.2.1 introduces our first proposed weighting approach, where consistency is ensured only by the household-level estimator. The estimator at the person level remains unaffected by the consistency requirements. As an estimator for the unknown totals of the common variables, we suggest a person-level estimator, because in household surveys the common variables might be primarily person characteristics, as mentioned in point 2 (given above). Section 4.2.2 presents our second proposed weighting approach, where we strive to improve the estimates of the unknown common variable totals. Thus, for each common variable total, a separate model is implemented. Consistent estimates between the person and the household level are guaranteed by incorporating the same estimated common variable totals into the estimators at both levels.

4.2.1 First Proposed Weighting Approach

In the first proposed weighting approach, consistency is ensured by incorporating the common variables in the household-level estimator. This subsection is organized as follows: The point estimators and corresponding weights of our first alternative weighting approach are deduced in Section 4.2.1.1. In this regard, we decompose the formulas for the estimator such that the effect caused by the consistency requirements can be quantified. In Section 4.2.1.2, we derive the analytical expression of the variance estimators that considers the additional source of randomness induced by the estimated common variable totals. Since the resulting variance formulas are cumbersome, Section 4.2.1.3 employs the computational proceeding.

Before we introduce our alternative weighting approaches, we initially present some basic notation for better readability of this chapter. The presentation here serves as an overview. The detailed notation is repeated when it is explicitly used in the respective formulas.

Auxiliary variables
Let $\boldsymbol{x_i} = (x_{i1}, \dots, x_{iq}, \dots, x_{iQ})^T$ be the Q- dimensional auxiliary vector of person i. To emphasize that different auxiliary variables can be included at the person and household level, we denote the auxiliary vector at the household level by an extra letter. Thus, consider $\boldsymbol{a_g} = (a_{g1}, \dots, a_{gk}, \dots, a_{gK})^T$ as the K dimensional auxiliary vector of household g. The known vectors of the totals are given by $\boldsymbol{T_x}$ and $\boldsymbol{T_a}$. The person-level common variables vector $\boldsymbol{c_i} = (c_{i1}, \dots, c_{il}, \dots, c_{iL})^T$ of dimension L sums up per household to $\sum_{i \in U_g} \boldsymbol{c_i} = \boldsymbol{c_g} = (c_{g1}, \dots, c_{gl}, \dots, c_{gL})^T$. The totals of the common variables are unknown and have to be estimated by $\boldsymbol{\tilde{T}_c}$. The GLS estimator is based on the combined person- and household-level auxiliary variables. To clearly differentiate between the level-specific and combined information, we use the Greek letters $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and $\boldsymbol{\kappa}$ for the combined auxiliaries. Their exact definitions are given in Section 4.3.

Estimators and Subscripts
To distinguish between the total estimators at the person and household level, we use different subscripts. $\hat{T}_{y_p}$ is the estimator at the person level and $\hat{T}_{y_h}$ is the estimator at the household level. However, we skip the subscript indicating the estimation level in $\boldsymbol{\hat{T}_x^{\text{HT}}}$ and $\boldsymbol{\hat{T}_a^{\text{HT}}}$,

since the auxiliaries $\boldsymbol{x_i}$ and $\boldsymbol{a_g}$ are defined only at one level. We omit also the subscript in the Horvitz-Thompson estimator for the common variables $\hat{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{HT}}$, to emphasize the equality of both estimators at the person and household-level.

Coefficients
For the decomposition of the estimators into a naïve GREG estimator and an adjustment term, we distinguish between various regression coefficients. The coefficients of the naïve GREG estimator are denoted by $\hat{\boldsymbol{B}}_{\boldsymbol{x}}$ or $\hat{\boldsymbol{B}}_{\boldsymbol{a}}$. The coefficients of auxiliary models to derive the adjustment term are defined as $\hat{\boldsymbol{F}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{F}}_{\boldsymbol{a}}$. For the coefficients of the estimators containing both the auxiliary variables and the common variables, we use different letters for the person level and the household level, since the common variables emerge at both levels. Thus, at the person level, the coefficients are described by $\hat{\boldsymbol{D}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{D}}_{\boldsymbol{c}}$. At the household level, the coefficients are presented by $\hat{\boldsymbol{E}}_{a}$ and $\hat{\boldsymbol{E}}_{c}$.

4.2.1.1 Point Estimation and Weights

The estimators of the first weighting approach are abbreviated with WA1. In contrast to integrated weighting, separate estimators are implemented at the person level and the household level. At the person level, the first proposed estimator is given by a naïve GREG estimator (as clarified in Definition 2)

$$\hat{T}_{y_p}^{\text{WA1}} = \hat{T}_{y_p}^{\text{GREG}} = \hat{T}_{y_p}^{\text{HT}} + \hat{\boldsymbol{B}}_{\boldsymbol{x}}^{T}(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}), \tag{4.1}$$

where $\hat{T}_{y_p}^{\text{HT}}$ is the Horvitz-Thompson estimator for the variable of interest. Estimated and known vectors of dimension Q of the auxiliary totals are denoted by $\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}$ and $\boldsymbol{T_x}$. The coefficient vector is expressed by

$$\hat{\boldsymbol{B}}_{\boldsymbol{x}} = (\sum_{i\in s_p} \pi_i^{-1}\boldsymbol{x_i}\boldsymbol{x_i}^T)^{-1}\sum_{i\in s_p}\pi_i^{-1}\boldsymbol{x_i}y_i.$$

Accordingly, the proposed estimator at the person level, $\hat{T}_{y_p}^{\text{WA1}}$, contains only the auxiliary variables $\boldsymbol{x_i}$ and thus remains unaffected by the consistency requirements. For simplicity, we assume a constant variance component of $v_i = 1$ (for a definition of the variance parameter see Section 2.3). The weights of $\hat{T}_{y_p}^{\text{WA1}}$ are then defined as

$$w_i^{\text{WA1}} = \pi_i^{-1} + \pi_i^{-1}\boldsymbol{x_i}^T(\sum_{i\in s_p}\pi_i^{-1}\boldsymbol{x_i}\boldsymbol{x_i}^T)^{-1}(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}). \tag{4.2}$$

At the household level, a separate estimator is implemented. To ensure consistency between person- and household-level estimates, we incorporate the common variables as additional auxiliaries. The totals of the common variables are unknown and have to be estimated by $\tilde{\boldsymbol{T}}_{\boldsymbol{c}}$. We skip the subscript indicating the level of estimation, because $\tilde{\boldsymbol{T}}_{\boldsymbol{c}}$ must be equal at the person and household-level. Because the common variables are initially person characteristics, we suggest estimating $\tilde{\boldsymbol{T}}_{\boldsymbol{c}}$ by a person-level estimator determined by

$$\tilde{\boldsymbol{T}}_{\boldsymbol{c}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{WA1}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} = \sum_{i\in s_p} w_i^{\text{WA1}}\boldsymbol{c_i}, \tag{4.3}$$

where w_i^{WA1} is obtained from (4.2). Hence, $\tilde{\boldsymbol{T}}_{\boldsymbol{c}}$ is estimated by the proposed person-level estimator.

The auxiliary vector at the household level is denoted by an extra letter, a_g, to emphasize that our alternative weighting approach allows inclusion of different auxiliaries at both levels. The known and estimated total vectors of dimension K are denoted by $\boldsymbol{T_a}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{a}}^{\text{HT}}$, respectively. Given (4.3), we propose the following estimator at the household level

$$\hat{T}_{y_h}^{\text{WA1}} = \hat{T}_{y_h}^{\text{HT}} + \hat{\boldsymbol{E}}_{\boldsymbol{a}}^{T}(\boldsymbol{T_a} - \hat{\boldsymbol{T}}_{\boldsymbol{a}}^{\text{HT}}) + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{HT}}), \tag{4.4}$$

where the coefficients $\hat{\boldsymbol{E}}_{\boldsymbol{a}}$ and $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ are simultaneously estimated by

$$\begin{pmatrix} \hat{\boldsymbol{E}}_{\boldsymbol{a}} \\ \hat{\boldsymbol{E}}_{\boldsymbol{c}} \end{pmatrix} = \left[\sum_{g \in s_h} \pi_g^{-1} \begin{pmatrix} \boldsymbol{a_g} \\ \boldsymbol{c_g} \end{pmatrix} \begin{pmatrix} \boldsymbol{a_g} \\ \boldsymbol{c_g} \end{pmatrix}^T \right]^{-1} \sum_{g \in s_h} \pi_g^{-1} \begin{pmatrix} \boldsymbol{a_g} \\ \boldsymbol{c_g} \end{pmatrix} y_g. \tag{4.5}$$

It is assumed that the partitioned matrix $\sum_{g \in s_h} \pi_g^{-1} \begin{pmatrix} \boldsymbol{a_g} \\ \boldsymbol{c_g} \end{pmatrix} \begin{pmatrix} \boldsymbol{a_g} \\ \boldsymbol{c_g} \end{pmatrix}^T$ is of full rank $K+L$. Subscript h refers to the household level.

Remark 2. *Even if the person- and household-level variables of interest and auxiliaries are equal, and thus the person-level variables sum up to the household-level variables, the GREG estimators for both the common variables and the variables of interest differ from each other:* $\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG} \neq \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}$ *and* $\hat{T}_{y_p}^{GREG} \neq \hat{T}_{y_h}^{GREG}$*. The inequality is caused by the divergent strength of the relationships between* $\boldsymbol{c_i}$ *or* y_i *and* $\boldsymbol{x_i}$ *as well as between* $\boldsymbol{c_g}$ *or* y_g *and* $\boldsymbol{x_g}$ *(see Section 3.2.2 for details).*

Now, we are interested in quantifying the impact of ensuring consistency on our proposed estimator (4.4). For this purpose, we decompose the household-level estimator into a naïve GREG estimator, which does not ensure consistent estimates, and an adjustment term capturing the effect caused by including the common variables as additional auxiliaries. For the decomposition of the estimator, an orthogonal decomposition of the coefficients (cf. Seber, 1977) is applied. The orthogonal decomposition is given by

$$\underset{(K\times 1)}{\hat{\boldsymbol{E}}_{\boldsymbol{a}}} = \underset{(K\times 1)}{\hat{\boldsymbol{B}}_{\boldsymbol{a}}} - \underset{(K\times L)}{\hat{\boldsymbol{F}}_{\boldsymbol{a}}} \underset{(L\times 1)}{\hat{\boldsymbol{E}}_{\boldsymbol{c}}}, \tag{4.6}$$

where $\hat{\boldsymbol{B}}_{\boldsymbol{a}}$ arises from

$$\hat{T}_{y_h}^{\text{GREG}} = \hat{T}_{y_h}^{\text{HT}} + \hat{\boldsymbol{B}}_{\boldsymbol{a}}^{T}(\boldsymbol{T_a} - \hat{\boldsymbol{T}}_{\boldsymbol{a}}^{\text{HT}}) \tag{4.7}$$

as a naïve GREG estimator at the household level solely containing $\boldsymbol{a_g}$ as auxiliaries. The product of $\hat{\boldsymbol{F}}_{\boldsymbol{a}}$ and $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ captures the effects of the common variables on the variable of interest neglected by $\hat{\boldsymbol{B}}_{\boldsymbol{a}}$. Coefficient matrix $\hat{\boldsymbol{F}}_{\boldsymbol{a}}$ is obtained from

$$\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{HT}} + \hat{\boldsymbol{F}}_{\boldsymbol{a}}^{T}(\boldsymbol{T_a} - \hat{\boldsymbol{T}}_{\boldsymbol{a}}^{\text{HT}})$$

as an L-dimensional vector comprising the household-level GREG estimators for the unknown common variable totals with $\boldsymbol{a_g}$ as auxiliaries. Hence, $\hat{\boldsymbol{F}}_{\boldsymbol{a}}$ describes the extent to which $\boldsymbol{a_g}$ helps to predict the common variables $\boldsymbol{c_g}$. Coefficient vector $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ is defined in (4.5) and describes the effect of $\boldsymbol{c_g}$ on the variable of interest y_g controlled for the effects of the auxiliaries $\boldsymbol{a_g}$.

Inserting the orthogonal decomposition (4.6) into (4.4), we obtain

$$\begin{aligned}
\hat{T}_{y_h}^{\mathrm{WA1}} &= \hat{T}_{y_h}^{\mathrm{HT}} + (\hat{\boldsymbol{B}}_{\boldsymbol{a}} - \hat{\boldsymbol{F}}_{\boldsymbol{a}}\hat{\boldsymbol{E}}_{\boldsymbol{c}})^T(\boldsymbol{T_a} - \hat{\boldsymbol{T}}_{\boldsymbol{a}}^{\mathrm{HT}}) + \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{HT}}) \\
&= \underbrace{\hat{T}_{y_h}^{\mathrm{HT}} + \hat{\boldsymbol{B}}_{\boldsymbol{a}}{}^T(\boldsymbol{T_a} - \hat{\boldsymbol{T}}_{\boldsymbol{a}}^{\mathrm{HT}})}_{\hat{T}_{y_h}^{\mathrm{GREG}}} - \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T\hat{\boldsymbol{F}}_{\boldsymbol{a}}{}^T(\boldsymbol{T_a} - \hat{\boldsymbol{T}}_{\boldsymbol{a}}^{\mathrm{HT}}) + \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{HT}}) \\
&= \hat{T}_{y_h}^{\mathrm{GREG}} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T\Big(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}} - \underbrace{\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{HT}} - \hat{\boldsymbol{F}}_{\boldsymbol{a}}{}^T(\boldsymbol{T_a} - \hat{\boldsymbol{T}}_{\boldsymbol{a}}^{\mathrm{HT}})}_{\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}}\Big) \\
&= \underbrace{\hat{T}_{y_h}^{\mathrm{GREG}}}_{a)} + \underbrace{\hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}})}_{b)}\,. \qquad (4.8)
\end{aligned}$$

Therefore, the first proposed estimator at the household level $\hat{T}_{y_h}^{\mathrm{WA1}}$ can be decomposed into:

a) a naïve GREG estimator defined in (4.7) omitting the common variables and

b) an adjustment term capturing the impact induced by the consistency requirements and thus incorporating the common variables into the estimator.

Adjustment term b) depends on the coefficient vector $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ defined in (4.5) and the difference between the person- and household-level estimates for the common variable totals. The greater the difference between the two estimators, the greater the adjustment term. The partial coefficient $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ can, alternatively to (4.5), be expressed in terms of residuals, given by

$$\hat{\boldsymbol{E}}_{\boldsymbol{c}} = \left(\sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}T}\right)^{-1}\sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}} r_g^{B_a} \qquad (4.9)$$

with $r_g^{B_a} = y_g - \hat{\boldsymbol{B}}_{\boldsymbol{a}}{}^T\boldsymbol{a_g}$ and $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}} = \boldsymbol{c_g} - \hat{\boldsymbol{F}}_{\boldsymbol{a}}{}^T\boldsymbol{a_g}$ resulting from regressing the variable of interest or common variables on the auxiliaries. The following remark helps to comprehend the notation of $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ in terms of residuals in (4.9).

Remark 3. ***Expression of Partial Coefficients in Terms of Residuals***
Define $\boldsymbol{A} = (\boldsymbol{a_1}^T, \ldots, \boldsymbol{a_g}^T, \ldots, \boldsymbol{a}_{\boldsymbol{m}}^T)$ as $m \times K$ auxiliary matrix and $\boldsymbol{y} = (y_1, \ldots, y_i, \ldots, y_n)^T$ as n-vector describing the variable of interest. Then, $\hat{\boldsymbol{E}}_{\boldsymbol{c}} = (\sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}T})^{-1}\sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{h}}^{\boldsymbol{F_a}} r_g^{B_a}$ is the vector pendant to the residual maker matrix representation of $\hat{\boldsymbol{E}}_{\boldsymbol{c}} = (\boldsymbol{A}^T\boldsymbol{M}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{M}\boldsymbol{y}$ known from econometric textbooks (cf. Greene, 2003, p. 244) with $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ as the idempotent residual maker matrix.

In accordance with Remark 3, the weights of our first proposed household-level estimator are determined by

$$w_g^{\text{WA1}} = w_g^{\text{GREG}} + \boldsymbol{r}_g^{\boldsymbol{F_a}\,T} \left(\sum_{g \in s_h} \boldsymbol{r}_g^{\boldsymbol{F_a}} \boldsymbol{r}_g^{\boldsymbol{F_a}\,T} \right)^{-1} (\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \tag{4.10}$$

with w_g^{GREG} obtained from the naïve household-level GREG defined in (4.7).

To conclude, in our first proposed weighting approach, the person-level estimator (4.1) is determined by a naïve GREG estimator and remains unaffected by the consistency requirements. Consistency between estimates of the common variables is ensured by inserting $\tilde{\boldsymbol{T}}_{\boldsymbol{c}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}$ into the household-level estimator (4.4) or (4.8). The impact of the consistency requirements on our first proposed household-level estimator is quantified by adjustment term b) in (4.8). Because only the household-level estimator is adjusted by the common variables, our first weighting approach considerably facilitates the application for statistical offices.

For clarity, the following result summarizes the currently introduced point estimators of the first alternative weighting approach.

Result 4. ***First Weighting Approach***
In our first alternative weighting approach, we suggest $\tilde{\boldsymbol{T}}_{\boldsymbol{c}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}$ *as the estimator of the unknown common variable totals. The person- and household-level estimators are given by*

$$\begin{aligned} \hat{T}_{y_p}^{WA1} &= \hat{T}_{y_p}^{GREG} \\ \hat{T}_{y_h}^{WA1} &= \hat{T}_{y_h}^{GREG} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T} (\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}) \end{aligned} \tag{4.11}$$

where $\hat{T}_{y_p}^{GREG}$ *and* $\hat{T}_{y_h}^{GREG}$ *are the naïve GREG estimators defined in (4.1) and (4.7).* $\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}$ *and* $\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}$ *are the person- and household-level GREG estimators for the common totals with* $\boldsymbol{x_i}$ *and* $\boldsymbol{a_g}$ *as auxiliaries, respectively. The household-level coefficient expressed in terms of residuals is defined as*

$$\hat{\boldsymbol{E}}_{\boldsymbol{c}} = \left(\sum_{g \in s_h} \boldsymbol{r}_g^{\boldsymbol{F_a}} \boldsymbol{r}_g^{\boldsymbol{F_a}\,T} \right)^{-1} \sum_{g \in s_h} \boldsymbol{r}_g^{\boldsymbol{F_a}} r_g^{B_a}$$

with $r_g^{B_a} = y_g - \hat{\boldsymbol{B}}_{\boldsymbol{a}}^{T} \boldsymbol{a_g}$ *and* $\boldsymbol{r}_g^{\boldsymbol{F_a}} = \boldsymbol{c_g} - \hat{\boldsymbol{F}}_{\boldsymbol{a}}^{T} \boldsymbol{a_g}$ *resulting from regressing the variable of interest or common variables on the auxiliaries.*

Proof. Given by Definition 2 and by inserting (4.6) into (4.4). □

4.2.1.2 Variance Estimation

In the following, we differentiate between the variance estimation for ordinary and common variables.

Variance Estimation for Ordinary Variables

At the person level, the variance estimator of our first proposed estimator (4.1) is determined by the variance of a naïve GREG estimator under cluster sampling

$$\hat{V}(\hat{T}_{y_p}^{\text{WA1}}) = \hat{V}(\hat{T}_{y_p}^{\text{GREG}}) = \sum_{g \in s_h} \sum_{k \in s_h} \frac{\triangle_{gk}}{\pi_{gk}} \sum_{i \in U_g} w_i^{\text{WA1}} r_i \sum_{j \in U_k} w_j^{\text{WA1}} r_j \tag{4.12}$$

with $\triangle_{gk} = \pi_{gk} - \pi_g \pi_k$, residuals $r_i = y_i - \boldsymbol{x_i}^T \hat{\boldsymbol{B}}_{\boldsymbol{x}}$ and w_i^{WA1} defined in (4.2).

At the household-level, the variance estimator should take into account the additional source of randomness induced by inserting the estimated common variables totals instead of known population totals. To respect the additional random source in the Taylor linearization, we have to supplementarily differentiate the random function with respect to the estimated common variable totals. We use expression (4.8) in the Taylor linearization instead of (4.4) because the former expression simplifies the derivatives.

Result 5. ***Variance Estimator of the First Household-Level Estimator***
The variance estimator of the first proposed household-level estimator $\hat{T}_{y_h}^{WA1}$ using the Taylor linearization technique is given by

$$\hat{V}(\hat{T}_{y_h}^{WA1}) \doteq \widehat{V}_1 + \widehat{V}_2 + \widehat{V}_3 + 2\widehat{V}_{12} - 2\widehat{V}_{13} - 2\widehat{V}_{23} \tag{4.13}$$

with

$$\begin{aligned}
\hat{V}_1 &= \hat{V}(\hat{T}_{y_h}^{GREG}), & \widehat{V}_{12} &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\, \widehat{Cov}(\hat{T}_{y_h}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}),\\
\hat{V}_2 &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\, \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}) \hat{\boldsymbol{E}}_{\boldsymbol{c}}, & \widehat{V}_{13} &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\, \widehat{Cov}(\hat{T}_{y_h}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}),\\
\hat{V}_3 &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\, \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}) \hat{\boldsymbol{E}}_{\boldsymbol{c}}, & \widehat{V}_{23} &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\, \widehat{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}) \hat{\boldsymbol{E}}_{\boldsymbol{c}},
\end{aligned}$$

where $\widehat{Cov}$ denotes the estimated covariance. Estimated variances and covariances can be obtained by (2.10) by inserting the appropriate variables.

Proof. We deduce the Taylor linearization according to Särndal et al. (1992, p. 173, p. 236). Equation (4.8) can be rewritten as function of random terms

$$\hat{T}_{y_h}^{\text{WA1}} = f(\hat{T}_{y_h}^{\text{GREG}}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{z}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}), \tag{4.14}$$

where the non-linear coefficient is decomposed into

$$\hat{\boldsymbol{E}}_{\boldsymbol{c}} = \left(\sum_{g \in s_h} \boldsymbol{r_g^{F_a}} \boldsymbol{r_g^{F_a}}^T \right)^{-1} \sum_{g \in s_h} \boldsymbol{r_g^{F_a}} r_g^{B_a} = \hat{\boldsymbol{Z}}^{-1} \hat{\boldsymbol{z}}$$

with

$$\hat{\boldsymbol{Z}} = \sum_{g \in s_h} \frac{\boldsymbol{r_g^{F_a}} \boldsymbol{r_g^{F_a}}^T}{\pi_g} \quad \text{as a } L \times L\text{-matrix with elements} \quad \hat{z}_{ll'} = \sum_{g \in s_h} \frac{r_{gl}^{F_a} r_{gl'}^{F_a}{}^T}{\pi_g} \quad \text{and}$$

$$\hat{\boldsymbol{z}} = \sum_{g \in s_h} \frac{\boldsymbol{r_g^{F_a}} r_g^{B_a}}{\pi_g} \quad \text{as a } L\text{-vector with elements} \quad \hat{z}_{l0} = \sum_{g \in s_h} \frac{r_{gl}^{F_a} r_g^{B_a}}{\pi_g}.$$

The residuals are obtained by $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}} = \boldsymbol{c_g} - \hat{\boldsymbol{F}}_{\boldsymbol{a}}^{T} \boldsymbol{a_g}$ and $r_g^{B_a} = y_g - \hat{\boldsymbol{B}}_{\boldsymbol{a}}^{T} \boldsymbol{a_g}$.

To approximate the non-linear function $f(\hat{T}_{y_h}^{\text{GREG}}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{z}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})$ by a linear function, we derive the partial derivatives with respect to all random terms. In $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ each element $\hat{z}_{ll'}$ and $\hat{z}_{l0}$ has to be differentiated with respect to $\boldsymbol{c_l}$ with $l = 1, \dots, L$. The following partial derivatives result

$$
\begin{aligned}
\frac{\partial f}{\partial \hat{T}_{y_h}^{\text{GREG}}} &= 1, \\
\frac{\partial f}{\partial \hat{T}_{c_{p,l}}^{\text{GREG}}} &= E_{c_l} \qquad \text{for all } \ l = 1, \dots, L, \\
\frac{\partial f}{\partial \hat{T}_{c_{h,l}}^{\text{GREG}}} &= -E_{c_l} \qquad \text{for all } \ l = 1, \dots, L, \\
\frac{\partial f}{\partial \hat{z}_{ll'}} &= (\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})^T (-\hat{\boldsymbol{Z}}^{-1} \frac{\partial \hat{\boldsymbol{Z}}}{\partial z_{ll'}} \hat{\boldsymbol{Z}}^{-1}) \hat{\boldsymbol{z}} \\
&= (\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})^T (-\hat{\boldsymbol{Z}}^{-1} \boldsymbol{\Lambda}_{ll'} \hat{\boldsymbol{Z}}^{-1}) \hat{\boldsymbol{z}} \qquad \text{for all } \ l, l' = 1, \dots, L \ \text{ and } \ l \leq l', \\
\frac{\partial f}{\partial \hat{z}_{l0}} &= (\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})^T \hat{\boldsymbol{Z}}^{-1} \boldsymbol{\lambda}_l \qquad \text{for all } \ l = 1, \dots, L,
\end{aligned}
$$

with $\boldsymbol{\Lambda}_{ll'}$ as $L \times L$ matrix with value 1 in position (l, l') and (l', l) and 0 elsewhere as well as $\boldsymbol{\lambda}_l$ as L-vector with value 1 in position l and 0 elsewhere.

By inserting the partial derivatives into the Taylor series (2.8) and evaluating these at the expected values, $E(\hat{T}_{y_h}^{\text{GREG}}) = T_{y_h}$, $E(\hat{\boldsymbol{Z}}) = \boldsymbol{Z}$, $E(\hat{\boldsymbol{z}}) = \boldsymbol{z}$, $E(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}) = \boldsymbol{T_c}$, $E(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) = \boldsymbol{T_c}$, the linearized household-level GREG estimator results in

$$
\begin{aligned}
\hat{T}_{y_h}^{\text{WA1}} &\doteq (T_{y_h} + \boldsymbol{E_c}^T (\boldsymbol{T_c} - \boldsymbol{T_c}) + 1 \cdot (\hat{T}_{y_h}^{\text{GREG}} - T_{y_p}) + \sum_{l=1}^{L} E_{c_l} (\hat{T}_{c_{p,l}}^{\text{GREG}} - T_{c_l}) \\
&\quad - \sum_{l=1}^{L} E_{c_l} (\hat{T}_{c_{h,l}}^{\text{GREG}} - T_{c_l}) \\
&= \hat{T}_{y_h}^{\text{GREG}} + \boldsymbol{E_c}^T (\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}).
\end{aligned}
$$

It should be remarked that the expected values of both $\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}$ equal $\boldsymbol{T_c}$.

The approximated design-based variance of the linearized first proposed estimator at the household level is given by

$$
\begin{aligned}
\mathrm{V}(\hat{T}_{y_h}^{\mathrm{WA1}}) &\doteq \mathrm{V}\Big(\hat{T}_{y_h}^{\mathrm{GREG}} + \boldsymbol{E_c}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}})\Big) \\
&= \mathrm{V}(\hat{T}_{y_h}^{\mathrm{GREG}}) + \mathrm{V}(\boldsymbol{E_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}} - \boldsymbol{E_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}) \\
&\qquad + 2\mathrm{Cov}(\hat{T}_{y_h}^{\mathrm{GREG}}, \boldsymbol{E_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}} - \boldsymbol{E_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}) \\
&= \underbrace{\mathrm{V}(\hat{T}_{y_h}^{\mathrm{GREG}})}_{V_1} + \underbrace{\boldsymbol{E_c}^T\mathrm{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}})\boldsymbol{E_c}}_{V_2} + \underbrace{\boldsymbol{E_c}^T\mathrm{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}})\boldsymbol{E_c}}_{V_3} \\
&\qquad - 2\,\underbrace{\boldsymbol{E_c}^T\mathrm{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}})\boldsymbol{E_c}}_{V_{23}} \\
&\qquad + 2\,\underbrace{\boldsymbol{E_c}^T\mathrm{Cov}(\hat{T}_{y_h}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}})}_{V_{12}} -2\,\underbrace{\boldsymbol{E_c}^T\mathrm{Cov}(\hat{T}_{y_h}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}})}_{V_{13}}
\end{aligned}
$$

with Cov as the approximate covariance matrix and V_1 as population parameter of $\hat{V}_1$. $\hat{V}(\hat{T}_{y_h}^{\mathrm{WA1}})$ results by estimating $\mathrm{V}(\hat{T}_{y_h}^{\mathrm{WA1}})$ from the sample s_h by the plug-in method. □

Result 5 shows that the efficiency of our first household-level estimator $\hat{T}_{y_h}^{\mathrm{WA1}}$ depends on the accuracy of the residual variance of the point estimates $\hat{T}_{y_h}^{\mathrm{GREG}}$, $\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}$ as well as on their covariances. The higher the correlation between $\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}$, the higher the precision gain of our first household-level estimator. Variance components $\hat{\mathrm{V}}_{12}$ and $\hat{\mathrm{V}}_{23}$ depend on the observed dependence between the person- and household-level data set. Compared to the variance of a naïve GREG estimator, given by $\hat{V}_1$, the complexity of the variance of the proposed estimator is increased by five additional variance components: $\hat{V}_2$, $\hat{V}_3$, $\hat{V}_{12}$, $\hat{V}_{23}$ and $\hat{V}_{23}$. We show in the next section that the additional computational effort is, however, limited.

Variance Estimation of the Common Variables

Note that the following formulas refer to the L common variables. We start with the person-level estimator. Substituting the variables of interest by the common variables in (4.1) yields

$$\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{WA1}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{HT}} + \hat{\boldsymbol{B}}_{\boldsymbol{x}}^{T}(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\mathrm{HT}}).$$

Its variance is simply estimated by

$$V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{WA1}}) = V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}}). \tag{4.15}$$

We continue with substituting the common variables into the household-level estimator (4.4), which results in

$$\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{WA1}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}), \tag{4.16}$$

where

$$\hat{\boldsymbol{E}}_{\boldsymbol{c}} = (\sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}T})^{-1} \sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{B_a}} = \mathrm{diag}(1,\ldots,1) = \boldsymbol{I_L} \tag{4.17}$$

is a coefficient matrix of dimension $L \times L$ with $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{B_a}} = \boldsymbol{c_g} - \hat{\boldsymbol{B}}_{\boldsymbol{a}}{}^T \boldsymbol{x_g}$ and $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}} = \boldsymbol{c_g} - \hat{\boldsymbol{F}}_{\boldsymbol{a}}{}^T \boldsymbol{x_g}$. Let $\boldsymbol{I_L}$ be an identity matrix of dimension L. It is evident that in case of common variables as variables of interest, $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{B_a}}$ and $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}$ refer to the same variable of interest. It follows that $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{B_a}} = \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}$ and thus $\hat{\boldsymbol{E}}_{\boldsymbol{c}} = \boldsymbol{I_L}$. The diagonal form of $\hat{\boldsymbol{E}}_{\boldsymbol{c}} = (\hat{\boldsymbol{E}}_{\boldsymbol{c_1}}, \dots, \hat{\boldsymbol{E}}_{\boldsymbol{c_l}}, \dots, \hat{\boldsymbol{E}}_{\boldsymbol{c_L}})$ is straight forward, because $\hat{\boldsymbol{E}}_{\boldsymbol{c_l}}$ arises from a regression model with the common variable c_{il} as variable of interest on the left-hand side and simultaneously as explanatory variable on the right-hand side of the regression. Thus, c_{il} is completely explained via c_{il}. Inserting $\hat{\boldsymbol{E}}_{\boldsymbol{c}} = \boldsymbol{I_L}$ into (4.16), we obtain

$$\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{WA1}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}.$$

Then, the variance estimator at the household level can be expressed by

$$V(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{WA1}}) = V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}). \tag{4.18}$$

Accordingly, if the variable to estimate is common to the person and household level, the variance estimators coincide: $V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{WA1}}) = V(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{WA1}})$.

4.2.1.3 Computational Proceeding to Calculate $\hat{\boldsymbol{V}}(\hat{\boldsymbol{T}}_{\boldsymbol{y_h}}^{\textbf{WA1}})$

Compared to the variance formula of a naïve GREG estimator, the computation of the first proposed household-level estimator (4.13) is more complex. However, in this section, we explore the additional computational effort to calculate $\hat{V}(\hat{T}_{y_h}^{\text{WA1}})$ and find it is limited to the computation of the residuals $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_x}}$ and $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}$. To show this, we derive the formulas for every single variance component in (4.13).

The first variance component $\hat{V}_1$ is determined by the variance of a naïve GREG estimator under single-stage cluster sampling

$$\hat{V}_1 = \hat{V}(\hat{T}_{y_h}^{\text{GREG}}) = \sum_{g \in s_h} \sum_{k \in s_h} \frac{\triangle_{gk}}{\pi_{gk}} w_g^{\text{WA1}} r_g^{B_a} w_k^{\text{WA1}} r_k^{B_a}$$

with residuals $r_g^{B_a} = y_g - \boldsymbol{x_g}^T \hat{\boldsymbol{B}}_{\boldsymbol{a}}$ and w_g^{WA1} defined in (4.10). Hence, the remaining variance components V_2,V_3, V_{12}, V_{13}, and V_{23} constitute the additional computational effort of $\hat{V}(\hat{T}_{y_h}^{\text{WA1}})$ compared to a naïve GREG estimator.

The second variance component $\hat{V}_2$ can be expressed by

$$\begin{aligned}
\underset{(1\times1)}{\hat{V}_2} &= \underset{(1\times L)}{\hat{\boldsymbol{E}}_{\boldsymbol{c}}}{}^T \underset{(L\times L)}{\hat{\boldsymbol{V}}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}})} \underset{(L\times1)}{\hat{\boldsymbol{E}}_{\boldsymbol{c}}} \\
&= \left(\hat{E}_{c1}, \dots, \hat{E}_{c_L}\right) \begin{pmatrix} \hat{V}(\hat{T}_{c_{p1}}^{\text{GREG}}) & & \\ \widehat{\text{Cov}}(\hat{T}_{c_{p1}}^{\text{GREG}}, \hat{T}_{c_{p2}}^{\text{GREG}}) & \ddots & \\ \vdots & \ddots & \\ \widehat{\text{Cov}}(\hat{T}_{c_{p1}}^{\text{GREG}}, \hat{T}_{c_{pL}}^{\text{GREG}}) \cdots & \widehat{\text{Cov}}(\hat{T}_{c_{pL}}^{\text{GREG}}, \hat{T}_{c_{pL-1}}^{\text{GREG}}) & \hat{V}(\hat{T}_{c_{pL}}^{\text{GREG}}) \end{pmatrix} \begin{pmatrix} \hat{E}_{c1} \\ \vdots \\ \hat{E}_{c_L} \end{pmatrix},
\end{aligned}$$

where

$$\hat{V}(\hat{T}^{\text{GREG}}_{c_{pl}}) = \sum_{g \in s_h} \sum_{k \in s_h} \frac{\triangle_{gk}}{\pi_{gk}} w_g^{\text{WA1}} r_g^{F_x^l} w_k^{\text{WA1}} r_k^{F_x^l} \quad \text{for all } \; l = 1, \ldots, L$$

and

$$\widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{c_{pl}}, \hat{T}^{\text{GREG}}_{c_{pt}}) = \sum_{g \in s_h} \sum_{k \in s_h} \frac{\triangle_{gk}}{\pi_{gk}} w_g^{\text{WA1}} r_g^{F_x^l} w_k^{\text{WA1}} r_k^{F_x^t} \quad \text{for all } \; t \neq l$$

with $r_g^{F_x^l} = \sum_{i \in U_g} r_i^{F_x^l} = \sum_{i \in U_g} (c_{i_l} - \boldsymbol{x_i}^T \hat{\boldsymbol{F}}_{\boldsymbol{x}}^{\boldsymbol{l}})$. Hence, the computation of V_2 additionally requires the calculation of the residual vector $\boldsymbol{r_g^{F_x}} = (r_g^{F_x^1}, \ldots, r_g^{F_x^l}, \ldots, r_g^{F_x^L})^T$ obtained from $\boldsymbol{r_g^{F_x}} = \sum_{i \in U_g} \boldsymbol{r_i^{F_x}} = \sum_{i \in U_g} (\boldsymbol{c_i} - \boldsymbol{F_x}^T \boldsymbol{x_i})$. It is important to note that $\boldsymbol{r_g^{F_x}}$ is independent from the variable of interest and has to be computed only once in each sample. The coefficient vector $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ is already available from the point estimator in (4.4).

Variance component $\hat{V}_3$ is defined analogously to $\hat{V}_2$. The only difference is that V_3 relates to the household level and thus requires computing the residual vector $\boldsymbol{r_g^{F_a}} = (r_g^{F_a^1}, \ldots, r_g^{F_a^l}, \ldots, r_g^{F_a^L})^T$ resulting from $\boldsymbol{r_g^{F_a}} = \boldsymbol{c_g} - \boldsymbol{F_a}^T \boldsymbol{x_g}$.

Variance component $\hat{V}_4$ is obtained by

$$\begin{aligned} \underset{(1\times 1)}{\hat{V}_4} &= \underset{(1\times L)}{\hat{\boldsymbol{E}}_{\boldsymbol{c}}}^T \underset{(L\times 1)}{\widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{y_h}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}})} \\ &= \begin{pmatrix} \hat{E}_{c1}, \ldots, \hat{E}_{c_L} \end{pmatrix} \begin{pmatrix} \widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{y_h}, \hat{T}^{\text{GREG}}_{c_{p1}}) \\ \vdots \\ \widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{y_h}, \hat{T}^{\text{GREG}}_{c_{pL}}) \end{pmatrix}, \end{aligned}$$

where

$$\widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{y_h}, \hat{T}^{\text{GREG}}_{c_{pl}}) = \sum_{g \in s_h} \sum_{k \in s_h} \frac{\triangle_{gk}}{\pi_{gk}} w_g^{\text{WA1}} r_g^{B_a} w_k^{\text{WA1}} r_k^{F_x^l} \quad \text{for all } \; l = 1, \ldots, L$$

with $r_g^{B_a}$ and $r_k^{F_x^l}$ as defined in $\hat{V}_1$ and $\hat{V}_2$. Thus, variance component $\hat{V}_4$ does not increase the computational effort since the residuals $r_g^{B_a}$ and $r_k^{F_x^l}$ are already available.

Variance component $\hat{V}_5$ is analogously defined as $\hat{V}_4$ but relates to the household level. It does not involve additional computations.

Finally, variance component $\hat{V}_6$ can be expressed by

$$\begin{aligned} \underset{(1\times 1)}{\hat{V}_6} &= \underset{(1\times L)}{\hat{\boldsymbol{E}}_{\boldsymbol{c}}}^T \underset{(L\times L)}{\widehat{\text{Cov}}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})} \underset{(L\times 1)}{\hat{\boldsymbol{E}}_{\boldsymbol{c}}} \\ &= \begin{pmatrix} \hat{E}_{c1}, \ldots, \hat{E}_{c_L} \end{pmatrix} \begin{pmatrix} \widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{c_{p1}}, \hat{T}^{\text{GREG}}_{c_{h1}}) & \ldots & \widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{c_{p1}}, \hat{T}^{\text{GREG}}_{c_{hL}}) \\ \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{c_{pL}}, \hat{T}^{\text{GREG}}_{c_{h1}}) & \ldots & \widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{c_{pL}}, \hat{T}^{\text{GREG}}_{c_{hL}}) \end{pmatrix} \begin{pmatrix} \hat{E}_{c1} \\ \vdots \\ \hat{E}_{c_L} \end{pmatrix}, \end{aligned}$$

where

$$\widehat{\text{Cov}}(\hat{T}^{\text{GREG}}_{c_{pl}}, \hat{T}^{\text{GREG}}_{c_{hl}}) = \sum_{g \in s_h} \sum_{k \in s_h} \frac{\triangle_{gk}}{\pi_{gk}} w_g^{\text{WA1}} r_g^{F_x^l} w_k^{\text{WA1}} r_k^{B_a} \quad \text{for all } l = 1, \ldots, L.$$

In summary, the additional effort to compute $\hat{V}(\hat{T}^{\text{WA1}}_{y_h})$ compared to the variance of a naïve GREG estimator, determined by $\hat{V}_1$, is limited to the calculation of the following L-dimensional residual vectors

- $\boldsymbol{r_g^{F_x}} = \sum_{i \in U_g} \boldsymbol{r_i^{F_x}} = \sum_{i \in U_g} (\boldsymbol{c_i} - \boldsymbol{F_x}^T \boldsymbol{x_i})$ and
- $\boldsymbol{r_g^{F_a}} = \boldsymbol{c_g} - \boldsymbol{F_a}^T \boldsymbol{x_g}$.

Both residual vectors are independent from the variable of interest and thus have to be calculated only once in each sample. Variance components $\hat{V}_2$, $\hat{V}_3$, $\hat{V}_{12}$, $\hat{V}_{23}$, and $\hat{V}_{23}$ are computable by an appropriate combination of these residuals. Hence, the additional computational effort depends on the number of variables that are required to be consistent.

It is important to note that the discussed additional computational effort refers only to the proposed household-level estimator. The person-level estimator, on the other side, remains unaffected and equals a naïve GREG estimator.

4.2.2 Second Proposed Weighting Approach

In our second proposed weighting approach, we strive to improve the estimates of the unknown common variable totals $\tilde{\boldsymbol{T}}_{\boldsymbol{c}}$. The underlying idea is that every common variable c_l, with $l = 1, \ldots, L$, can be modeled by a separate set of specialized auxiliary variables $\boldsymbol{z_l}$. To ensure consistency between person- and household-level estimates, we insert the same $\tilde{\boldsymbol{T}}_{\boldsymbol{c}}$ into our proposed estimators at both levels. Point estimators and corresponding weights are introduced in Section 4.2.2.1. Section 4.2.2.2 presents the variance estimators of our second proposed alternative weighting approach.

4.2.2.1 Point Estimation and Weights

Let $\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p^*}} = (\hat{T}^{\text{GREG}}_{c^*_{p,1}}, \ldots, \hat{T}^{\text{GREG}}_{c^*_{p,l}}, \ldots, \hat{T}^{\text{GREG}}_{c^*_{p,L}})^T$ be the L-vector of estimates for the common variable totals, where

$$\hat{\boldsymbol{T}}^{\text{GREG}}_{c^*_{p,l}} = \hat{\boldsymbol{T}}^{\text{HT}}_{c_{p,l}} + \hat{\boldsymbol{B}}_{z_l}(\boldsymbol{T}_{z_l} - \hat{\boldsymbol{T}}^{\text{HT}}_{z_l})$$

is estimated by $\boldsymbol{z_l}$ and with the estimated coefficient $\hat{\boldsymbol{B}}_{\boldsymbol{z_l}} = (\sum_{i \in s_p} \pi_i^{-1} \boldsymbol{z}_{il} \boldsymbol{z}_{il}^T)^{-1} \sum_{i \in s_p} \pi_i^{-1} \boldsymbol{z}_{il} \boldsymbol{c}_{il}$. The auxiliary variable set $\boldsymbol{z_l}$ can be chosen for each l common variable separately. That could be, for example, the auxiliary variable set with the highest explanatory power for the respective common variable. The specialized auxiliary variables $\boldsymbol{z_l}$ may contain some of the auxiliaries

$\boldsymbol{x}_i$, but can also contain further auxiliaries with known totals. Our intention of separate modeling is to use the best available estimate for $\tilde{\boldsymbol{T}}_{\boldsymbol{c}}$. The second alternative weighting approach is abbreviated with WA2. Inserting $\tilde{\boldsymbol{T}}_{\boldsymbol{c}} = \hat{\boldsymbol{T}}_{\boldsymbol{c}_p^*}^{\text{GREG}}$ as estimator for the unknown common variable totals yields our second person-level GREG estimator

$$\hat{T}_{y_p}^{\text{WA2}} = \hat{T}_{y_p}^{\text{HT}} + \hat{\boldsymbol{D}}_{\boldsymbol{x}}^{T}(\boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}) + \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_p^*}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c}_p}^{\text{HT}}), \tag{4.19}$$

where the coefficients $\hat{\boldsymbol{D}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{D}}_{\boldsymbol{c}}$ are simultaneously estimated by

$$\begin{pmatrix} \hat{\boldsymbol{D}}_{\boldsymbol{x}} \\ \hat{\boldsymbol{D}}_{\boldsymbol{c}} \end{pmatrix} = \left[\sum_{i \in s_p} \pi_i^{-1} \begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{c}_i \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{c}_i \end{pmatrix}^T \right]^{-1} \sum_{i \in s_p} \pi_i^{-1} \begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{c}_i \end{pmatrix} y_i. \tag{4.20}$$

It is assumed that the partitioned matrix $\sum_{i \in s_p} \pi_i^{-1} \begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{c}_i \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{c}_i \end{pmatrix}^T$ is of full rank $Q + L$.

To quantify the impact of the consistency requirements, we decompose $\hat{T}_{y_p}^{\text{WA2}}$ into a naïve GREG estimator and an adjustment term capturing the effect caused by ensuring consistency. Analogously to the proceeding of the first weighting approach, an orthogonal decomposition (cf. Seber, 1977) is applied to decompose $\hat{\boldsymbol{D}}_{\boldsymbol{x}}$ into

$$\underset{(Q \times 1)}{\hat{\boldsymbol{D}}_{\boldsymbol{x}}} = \underset{(Q \times 1)}{\hat{\boldsymbol{B}}_{\boldsymbol{x}}} - \underset{(Q \times L)}{\hat{\boldsymbol{F}}_{\boldsymbol{x}}} \underset{(L \times 1)}{\hat{\boldsymbol{D}}_{\boldsymbol{c}}} \tag{4.21}$$

where $\hat{\boldsymbol{B}}_{\boldsymbol{x}}$ arises from

$$\hat{T}_{y_p}^{\text{GREG}} = \hat{T}_{y_p}^{\text{HT}} + \hat{\boldsymbol{B}}_{\boldsymbol{x}}^{T}(\boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}) \tag{4.22}$$

as a naïve GREG estimator at the person level containing only $\boldsymbol{x}_i$ as auxiliaries. The product of $\hat{\boldsymbol{F}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{D}}_{\boldsymbol{c}}$ captures the effects of the common variables on the variable of interest omitted by $\hat{\boldsymbol{B}}_{\boldsymbol{x}}$. Coefficient matrix $\hat{\boldsymbol{F}}_{\boldsymbol{x}}$ is obtained from

$$\hat{\boldsymbol{T}}_{\boldsymbol{c}_p}^{\text{GREG}} = \hat{\boldsymbol{T}}_{\boldsymbol{c}_p}^{\text{HT}} + \hat{\boldsymbol{F}}_{\boldsymbol{x}}^{T}(\boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}),$$

a vector containing the person-level GREG estimators for the common variable totals with $\boldsymbol{x}_i$ as auxiliaries. Hence, $\hat{\boldsymbol{F}}_{\boldsymbol{x}}$ describes the extent to which $\boldsymbol{x}_i$ helps to predict the common variables $\boldsymbol{c}_i$. The coefficient vector $\hat{\boldsymbol{D}}_{\boldsymbol{c}}$ is already defined in (4.20) and describes the effect of $\boldsymbol{x}_i$ on y_i controlled for the effects of $\boldsymbol{c}_i$.

By inserting the orthogonal decomposition (4.21) into (4.19), we obtain

$$\begin{aligned} \hat{T}_{y_p}^{\text{WA2}} &= \hat{T}_{y_p}^{\text{HT}} + (\hat{\boldsymbol{B}}_{\boldsymbol{x}} - \hat{\boldsymbol{F}}_{\boldsymbol{x}}\hat{\boldsymbol{D}}_{\boldsymbol{c}})^T(\boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}) + \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_p^*}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{HT}}) \\ &= \underbrace{\hat{T}_{y_p}^{\text{HT}} + \hat{\boldsymbol{B}}_{\boldsymbol{x}}^{T}(\boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})}_{\hat{T}_{y_p}^{\text{GREG}}} + \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}\Big(\hat{\boldsymbol{T}}_{\boldsymbol{c}_p^*}^{\text{GREG}} - \underbrace{\hat{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{HT}} - \hat{\boldsymbol{F}}_{\boldsymbol{x}}^{T}(\boldsymbol{T}_{\boldsymbol{x}} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})}_{\hat{\boldsymbol{T}}_{\boldsymbol{c}_p}^{\text{GREG}}}\Big) \\ &= \underbrace{\hat{T}_{y_p}^{\text{GREG}}}_{a)} + \underbrace{\hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_p^*}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c}_p}^{\text{GREG}})}_{b)}. \end{aligned} \tag{4.23}$$

Therefore, our second proposed estimator $\hat{T}_{y_p}^{\text{WA2}}$ can be decomposed into:

a) a naïve GREG estimator for the variable of interest omitting the common variables and

b) an adjustment term capturing the impact induced by the consistency requirements.

Adjustment term b) depends on the difference between the estimated common variable totals at the person and household level and on the coefficient vector $\hat{\boldsymbol{D}}_{\boldsymbol{c}}$ defined in (4.20). According to Remark 3, $\hat{\boldsymbol{D}}_{\boldsymbol{c}}$ can alternatively be expressed in terms of residuals, given by

$$\hat{\boldsymbol{D}}_{\boldsymbol{c}} = \left(\sum_{i \in s_p} \boldsymbol{r}_i^{\boldsymbol{F_x}} \boldsymbol{r}_i^{\boldsymbol{F_x}T} \right)^{-1} \sum_{i \in s_p} \boldsymbol{r}_i^{\boldsymbol{F_x}} r_i^{B_x} \tag{4.24}$$

with $r_i^{B_x} = y_i - \hat{\boldsymbol{B}}_{\boldsymbol{x}}^T \boldsymbol{x}_{\boldsymbol{i}}$ and $\boldsymbol{r}_i^{\boldsymbol{F_x}} = \boldsymbol{c}_{\boldsymbol{i}} - \hat{\boldsymbol{F}}_{\boldsymbol{x}}^T \boldsymbol{x}_{\boldsymbol{i}}$ resulting from regressing the variable of interest or the common variables on the auxiliaries, respectively.

At the household level, a separate estimator is implemented. To ensure consistent person- and household-level estimates, we insert the same $\hat{\boldsymbol{T}}_{\boldsymbol{c}_p^*}$ into the household-level estimator as used for the person-level estimator (4.19). Then, our second proposed estimator at the household level is obtained by

$$\hat{T}_{y_h}^{\text{WA2}} = \hat{T}_{y_h}^{\text{GREG}} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^T (\hat{\boldsymbol{T}}_{\boldsymbol{c}_p^*}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c}_h}^{\text{GREG}}) \tag{4.25}$$

with $\hat{\boldsymbol{E}}_{\boldsymbol{c}}$ already defined in (4.9). We refrain from a detailed derivation of formula (4.25), since it is analogously deduced as (4.23) and (4.8).

We learn from (4.23) and (4.25) that the higher the difference between the estimated common variable totals, utilizing a specialized auxiliary sets $\boldsymbol{z}_{\boldsymbol{il}}$, compared to $\boldsymbol{x}_{\boldsymbol{i}}$ or $\boldsymbol{a}_{\boldsymbol{g}}$, the higher the adjustment term.

Comparing our first and second weighting approach, it becomes apparent that the point estimators differ with respect to the implementation expense and the quality of the estimated unknown common variable totals. The implementation expense of our first weighting approach is lower, because to ensure consistency only the household-level estimator is adjusted by the common variables. As estimator for the unknown totals our proposed estimator at the person level is applied. The implementation expense of our second alternative weighting approach, on the other side, is more demanding, because both the person- and household-level estimators are affected by the consistency requirements. Moreover, the estimation of $\tilde{\boldsymbol{T}}_{\boldsymbol{c}} = \hat{\boldsymbol{T}}_{\boldsymbol{c}_p^*}^{\text{GREG}}$ increases the computational effort compared with the first weighting approach because for every single common variable the specialized auxiliary variables $\boldsymbol{z}_{\boldsymbol{l}}$ has to be determined. However, we expect a precision gain for the estimates of the common variables and for all variables correlated with the common variables. As a result, the choice between our first and second weighting approach is determined by a trade-off between the implementation expense and the quality of the final estimates.

For the special case of $\boldsymbol{z}_{\boldsymbol{il}} = \boldsymbol{z}_{\boldsymbol{i}} = \boldsymbol{x}_{\boldsymbol{i}}$ for all $l = 1, \ldots, L$, our first weighting approach coincide with our second weighting approach, that is $\hat{T}_{y_p}^{\text{WA1}} = \hat{T}_{y_p}^{\text{WA2}}$ as well as $\hat{T}_{y_h}^{\text{WA1}} = \hat{T}_{y_h}^{\text{WA2}}$.

The weights of our second alternative weighting approach are given by

$$w_i^{\text{WA2}} = w_i^{\text{GREG}} + \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}^T}(\sum_{i\in s_p} \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}}\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}^T})^{-1}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}})$$

and

$$w_g^{\text{WA2}} = w_g^{\text{GREG}} + \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}^T}(\sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}^T})^{-1}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}).$$

For the sake of clarity, the following result summarizes our second weighting approach.

Result 6. ***Second Alternative Weighting Approach***
Let $\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{GREG} = (\hat{T}_{c_{p,1}^{*}}^{GREG}, \ldots, \hat{T}_{c_{p,l}^{*}}^{GREG}, \ldots, \hat{T}_{c_{p,L}^{*}}^{GREG})^T$ be the L-vector of estimates for the common variable totals, where $\hat{T}_{c_{p,l}^{*}}^{GREG}$ is estimated by $\boldsymbol{z_l}$, a specialized auxiliary variable set with the highest explanatory power for c_l, with $l = 1, \ldots, L$. The estimators of our second proposed alternative weighting approaches are then obtained from

$$\begin{aligned} \hat{T}_{y_p}^{WA2} &= \hat{T}_{y_p}^{GREG} + \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{GREG} - \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}) \quad \text{and} \\ \hat{T}_{y_h}^{WA2} &= \hat{T}_{y_h}^{GREG} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{GREG} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}), \end{aligned} \tag{4.26}$$

where $\hat{T}_{y_p}^{GREG}$ and $\hat{T}_{y_h}^{GREG}$ are naïve GREG estimators defined in (4.1) and (4.7). $\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}$ are the person- and the household-level GREG estimators for the common totals with $\boldsymbol{x_i}$ or $\boldsymbol{a_g}$ as auxiliaries, respectively. The person-level coefficient is defined by

$$\hat{\boldsymbol{D}}_{\boldsymbol{c}} = \left(\sum_{i\in s_p} \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}}\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}^T}\right)^{-1} \sum_{i\in s_p} \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}} r_i^{B_x},$$

where $r_i^{B_x} = y_i - \hat{\boldsymbol{B}}_{\boldsymbol{x}}^{T}\boldsymbol{x_i}$ results from regressing the variable of interest on the auxiliaries and $\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}} = \boldsymbol{c_i} - \hat{\boldsymbol{F}}_{\boldsymbol{x}}^{T}\boldsymbol{x_i}$. The household-level coefficient can be expressed as

$$\hat{\boldsymbol{E}}_{\boldsymbol{c}} = \left(\sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}^T}\right)^{-1} \sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}} r_g^{B_a},$$

where $r_g^{B_a} = y_g - \hat{\boldsymbol{B}}_{\boldsymbol{a}}^{T}\boldsymbol{a_g}$ and $\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}} = \boldsymbol{c_g} - \hat{\boldsymbol{F}}_{\boldsymbol{a}}^{T}\boldsymbol{a_g}$ result from regressing the variable of interest or common variables on the auxiliaries.

4.2.2.2 Variance Estimation

Analogously to the first weighting approach, we differentiate between the variance estimation for ordinary and common variables.

Variance Estimation for Ordinary Variables

In our second weighting approach, both the person- and household-level estimators contain the estimated common variable totals. Therefore, both variance estimators take the additional source of randomness into account. We refrain from deriving the variance estimators at this point and present only the results, since they are analogously deduced as for our first weighting approach, presented in Section 4.2.1.2. The interested reader is referred to Section B.1 in Appendix B. The variance estimators of the second estimators using the Taylor linearization technique are given by

$$\hat{V}(\hat{T}_{yt}^{\mathrm{WA2}}) \doteq \hat{V}_1 + \hat{V}_2 + \hat{V}_3 + 2\widehat{\mathrm{V}}_{12} - 2\widehat{\mathrm{V}}_{13} - 2\widehat{\mathrm{V}}_{23} \quad \text{for } t = \{p, h\}. \tag{4.27}$$

At the person level, the variance components are given by

$$\begin{aligned}
\hat{V}_1 &= \hat{V}(\hat{T}_{y_p}^{\mathrm{GREG}}), & \widehat{\mathrm{V}}_{12} &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}\widehat{\mathrm{Cov}}(\hat{T}_{y_p}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{\boldsymbol{GREG}}), \\
\hat{V}_2 &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}\hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{\mathrm{GREG}})\hat{\boldsymbol{D}}_{\boldsymbol{c}}, & \widehat{\mathrm{V}}_{13} &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}\widehat{\mathrm{Cov}}(\hat{T}_{y_p}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{\mathrm{GREG}}), \\
\hat{V}_3 &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}\hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{\mathrm{GREG}})\hat{\boldsymbol{D}}_{\boldsymbol{c}}, & \widehat{\mathrm{V}}_{23} &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}\widehat{\mathrm{Cov}}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{\mathrm{GREG}})\hat{\boldsymbol{D}}_{\boldsymbol{c}}.
\end{aligned} \tag{4.28}$$

At the household level, the variance components are obtained from

$$\begin{aligned}
\hat{V}_1 &= \hat{V}(\hat{T}_{y_h}^{\mathrm{GREG}}), & \widehat{\mathrm{V}}_{12} &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\widehat{\mathrm{Cov}}(\hat{T}_{y_h}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{\mathrm{GREG}}), \\
\hat{V}_2 &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{\mathrm{GREG}})\hat{\boldsymbol{E}}_{\boldsymbol{c}}, & \widehat{\mathrm{V}}_{13} &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\widehat{\mathrm{Cov}}(\hat{T}_{y_h}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{h}}}^{\mathrm{GREG}}), \\
\hat{V}_3 &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{h}}}^{\mathrm{GREG}})\hat{\boldsymbol{E}}_{\boldsymbol{c}}, & \widehat{\mathrm{V}}_{23} &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\widehat{\mathrm{Cov}}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{h}}}^{\mathrm{GREG}})\hat{\boldsymbol{E}}_{\boldsymbol{c}}.
\end{aligned} \tag{4.29}$$

$\widehat{\mathrm{Cov}}$ denotes the estimated covariance. Estimated variances and covariances can be obtained by (2.10) by inserting the appropriate variables.

The variance components of the person-level estimator (4.28) depend solely on the person level, whereas the variance components in (4.29) are influenced by person- and household-level estimates. The computational proceeding to calculate (4.28) and (4.29) is analogously given as in Section 4.2.1.3 by inserting the appropriate estimators.

Compared to our first weighting approach, the computation of the variance estimators of our second alternative weighting approach is more demanding. The additional computational effort confirms with the trade-off, as mentioned for the point estimator, between the implementation expense and the quality of the final estimates.

Variance Estimation for the Common Variables

Inserting the common variables into our second proposed estimator (4.19), the following person-level estimator results

$$\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{\mathrm{WA2}} = \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{\mathrm{GREG}} + \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^{*}}^{\mathrm{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{\mathrm{GREG}}),$$

where the $(L \times L)$ coefficient matrix is given by

$$\hat{\boldsymbol{D}}_{\boldsymbol{c}} = \left(\sum_{i \in s_p} \boldsymbol{r}_i^{\boldsymbol{F_x}} \boldsymbol{r}_i^{\boldsymbol{F_x}\,T} \right)^{-1} \sum_{i \in s_p} \boldsymbol{r}_i^{\boldsymbol{F_x}} \boldsymbol{r}_i^{\boldsymbol{B_x}} = \text{diag}(1, \dots, 1) = \boldsymbol{I_L}.$$

The residuals $\boldsymbol{r}_i^{\boldsymbol{B_x}} = \boldsymbol{c}_i - \hat{\boldsymbol{B}}_{\boldsymbol{x}}^{\;T} \boldsymbol{x}_i$ and $\boldsymbol{r}_i^{\boldsymbol{F_x}} = \boldsymbol{c}_i - \hat{\boldsymbol{F}}_{\boldsymbol{x}}^{\;T} \boldsymbol{x}_i$ are equal, because both refer to the same variable of interest. The diagonal form of $\hat{\boldsymbol{D}}_{\boldsymbol{c}} = (\hat{\boldsymbol{D}}_{\boldsymbol{c}_1}, \dots, \hat{\boldsymbol{D}}_{\boldsymbol{c}_l}, \dots, \hat{\boldsymbol{D}}_{\boldsymbol{c}_L}) = \text{diag}(1, \dots, 1)$ follows, since c_{il} is completely explained via c_{il}, as argued in detail in Section 4.2.1.2. Hence, $\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{WA2}} = \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}$. Then the variance estimator of the second person-level estimator is given by

$$V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{WA2}}) = V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\text{GREG}}). \tag{4.30}$$

The second household-level estimator (4.25) for the common variables as variables of interest can be expressed by

$$\hat{T}_{c_h}^{\text{WA2}} = \hat{T}_{c_h}^{\text{GREG}} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{\;T} (\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}),$$

where $\hat{\boldsymbol{E}}_{\boldsymbol{c}} = \text{diag}(1, \dots, 1)^T$ for the same reasons as given above. The variance estimator is obtained from

$$V(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{WA2}}) = V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\text{GREG}}). \tag{4.31}$$

Accordingly, it is valid that the variance estimators at the person level and the household level coincide, $V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{WA2}}) = V(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{WA2}})$, as in the first proposed weighting approach.

4.2.3 Distinction between the Alternative Weighting Approaches and the Method of Renssen and Niewenbroek (1993)

In Section 4.2, we discussed extensively the differences between our alternative weighting approaches and the method proposed by Renssen and Nieuwenbroek (1997) including the dependency between the surveys at hand, the definition of the common variables and the target populations. This section aims to derive how our proposed point and variance estimators differ from the point and variance estimators of Renssen and Nieuwenbroek (1997) due to these differences. For this purpose, we briefly review their suggested point and variance estimators. At this point, we focus more on discussing the differences than on deriving the formulas in detail. Therefore, we refer the interested reader to Section B.2 in Appendix B for a detailed derivation.

We start with the point estimators. Adopting the method of Renssen and Nieuwenbroek (1997, p. 371) to household surveys yields the following person- and household-level estimators

$$\begin{aligned} \hat{T}_{y_p}^{\text{RN}} &= \hat{T}_{y_p}^{\text{GREG}} + \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{\;T} (\tilde{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{RN}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}) \quad \text{and} \\ \hat{T}_{y_h}^{\text{RN}} &= \hat{T}_{y_h}^{\text{GREG}} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{\;T} (\tilde{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{RN}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}), \end{aligned} \tag{4.32}$$

where $\hat{T}_{y_p}^{\text{GREG}}$, $\hat{T}_{y_h}^{\text{GREG}}$, $\hat{\boldsymbol{D}}_c$ and $\hat{\boldsymbol{E}}_c$ are defined as in (4.24) and (4.9), respectively. RN indicates Renssen and Nieuwenbrok.

For the estimation of the unknown population total $\tilde{\boldsymbol{T}}_c$ a composite estimator based on the weighted average of the single estimates obtained from each of the independent surveys is applied

$$\tilde{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{RN}} = \boldsymbol{Q}\hat{\boldsymbol{T}}_{\boldsymbol{c}_p}^{\text{GREG}} + (\boldsymbol{1} - \boldsymbol{Q})\hat{\boldsymbol{T}}_{\boldsymbol{c}_h}^{\text{GREG}},$$

where $\boldsymbol{Q}$ is a weighting matrix of dimension $(L \times L)$ with $\boldsymbol{Q} + (\boldsymbol{1} - \boldsymbol{Q}) = \boldsymbol{I}$. Different choices for $\boldsymbol{Q}$ are discussed. Inserting $\tilde{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{RN}}$ into (4.32), we obtain

$$\begin{aligned} \hat{T}_{y_p}^{\text{RN}} &= \hat{T}_{y_p}^{\text{GREG}} - \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}(\boldsymbol{1} - \boldsymbol{Q})(\hat{\boldsymbol{T}}_{\boldsymbol{c}_p}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c}_h}^{\text{GREG}}) \quad \text{and} \\ \hat{T}_{y_h}^{\text{RN}} &= \hat{T}_{y_h}^{\text{GREG}} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\boldsymbol{Q}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_p}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c}_h}^{\text{GREG}}). \end{aligned} \tag{4.33}$$

Therefore, their suggested point estimators differ from our proposed point estimators (4.11) or (4.26) in two respects. First, Renssen and Nieuwenbroek (1997) suggested a composite estimator for the unknown common variable totals requiring both the person-level and the household-level estimates. Moreover, their composite estimator entails the computation of the weighting matrix $\boldsymbol{Q}$. In contrast, we suggest a person-level estimator for the unknown common variable totals. Our choice is justified by the fact that in household surveys it is more prevalent that the common variables are initial person characteristics, which are supplementarily assigned in aggregated form to the household-level data set. Because of this, it is questionable to what extent their composite estimator is appropriate for estimating the unknown person-level total. This difference between the estimated common variable totals of our proposed weighting approaches and the approach by Renssen and Nieuwenbroek (1997) is mainly driven by the different definitions of common variables.

Second, both estimators (4.33) request the same auxiliary information, as can be seen from equation (8) in Renssen and Nieuwenbroek (1997, p. 371) or from the fact that otherwise inserting $\tilde{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{RN}}$ into (4.32) does not result in (4.33). In contrast, our alternative weighting approaches allow us to include different auxiliaries at both levels and are thus more flexible in their variable selection. Hence, our alternative weighting approaches take into account that the estimators refer to differing target populations.

There are also differences with respect to the variance estimators. Comparing the variance estimators of (4.33) (given in Section B.2 in Appendix B) to our proposed variances estimators, presented in (4.12), (4.13), (4.28) and (4.29), it becomes apparent that the main difference arises from the covariances between the estimators of the variables of interest, and the common variables, denoted by the covariance terms $\widehat{\text{Cov}}_{12}$ and $\widehat{\text{Cov}}_{13}$. These covariances are essential in capturing the dependence between the person and the household data set. Therefore, the differences between the variance estimators are driven by the dependence of the person and household data sets, the weighting matrix $\boldsymbol{Q}$ and the signs of the variance components.

4.3 GLS Estimator as a Benchmark Estimator

As presented in Section 4.1, Zieschang (1986, 1990) explored a GLS adjustment algorithm to combine the information from two independent surveys. It can also be adopted to ensure consistent estimates between person- and household-level estimates. Therefore, this section discusses the GLS adjustment algorithm as a benchmark estimator for our proposed alternative weighting approaches. Since the GLS adjustment algorithm is equivalent to a calibration estimator with a chi-squared distance function (see Section 2.3.4), we call it GLS estimator hereinafter.

This section is organized as follows: Section 4.3.1.1 discusses the point estimators of the GLS estimator. Furthermore, to conform the GLS estimator to our proposed alternative weighting approaches, we embed the GLS estimator into the GREG estimation framework. This is a promising exercise since having the same expressions allows us to directly compare our proposed weighting approaches to the GLS estimator. Section 4.3.1.2 derives the variance estimator of the GLS estimator. Finally, in Section 4.3.2, we briefly review the modified GLS estimator established by Merkouris (2004) to account for the effective sample sizes of the independent surveys.

4.3.1 GLS Estimator According to Zieschang (1986, 1990)

4.3.1.1 Point Estimation and Weights

We decided to initially present the GLS estimator in matrix notation, as originally introduced by Zieschang (1986, 1990), because the intension of the estimator is more comprehensible in matrix notation. Subsequently, we rewrite the matrix in vector notation to conform to the vector notation of our proposed alternative estimators. Consider

$$\underset{(n\times Q)}{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{x_1}^T \\ \vdots \\ \boldsymbol{x_i}^T \\ \vdots \\ \boldsymbol{x_n}^T \end{pmatrix} \quad \text{and} \quad \underset{(m\times K)}{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{a_1}^T \\ \vdots \\ \boldsymbol{a_g}^T \\ \vdots \\ \boldsymbol{a_m}^T \end{pmatrix}$$

as auxiliary matrices at the person and household level with $\boldsymbol{x_i} = (x_{i1}, \dots, x_{iL})^T$ and $\boldsymbol{a_g} = (a_{g1}, \dots, a_{gL})^T$ as defined in the previous section. The corresponding vectors of the known totals are given by $\boldsymbol{T_x}$ and $\boldsymbol{T_a}$. The matrices for the common variables at the person and household level are denoted as

$$\underset{(n\times L)}{\boldsymbol{C_p}} = \begin{pmatrix} \boldsymbol{c_1}^T \\ \vdots \\ \boldsymbol{c_i}^T \\ \vdots \\ \boldsymbol{c_n}^T \end{pmatrix} \quad \text{and} \quad \underset{(m\times L)}{\boldsymbol{C_h}} = \begin{pmatrix} \boldsymbol{c_1}^T \\ \vdots \\ \boldsymbol{c_g}^T \\ \vdots \\ \boldsymbol{c_m}^T \end{pmatrix}.$$

To pool the person- and household-level information, Zieschang (1986, 1990) combined the auxiliary matrices to one single matrix $\boldsymbol{Z}$, obtained by

$$\underset{(n+m)\times(Q+K+L)}{\boldsymbol{Z}} = \begin{pmatrix} \boldsymbol{X} & 0 & \boldsymbol{C_p} \\ 0 & \boldsymbol{A} & -\boldsymbol{C_h} \end{pmatrix}.$$

The combined total vector of dimension $(Q+K+L)$ is defined as $\boldsymbol{T_Z} = (\boldsymbol{T_x}^T, \boldsymbol{T_a}^T, \boldsymbol{0}^T)^T$. Suppose $\boldsymbol{\Pi} = \text{diag}(\boldsymbol{\Pi_p}, \boldsymbol{\Pi_h})$ as combined $(n+m)\times(n+m)$-weight matrix with submatrices $\boldsymbol{\Pi_p} = \text{diag}(\pi_1, \ldots, \pi_n)$ and $\boldsymbol{\Pi_h} = \text{diag}(\pi_1, \ldots, \pi_m)$. The combined $(n+m)$-vector of design weights is denoted by $\boldsymbol{d} = (\boldsymbol{d_p}, \boldsymbol{d_h})^T$ with $\boldsymbol{d_p} = (d_1, \ldots, d_n)$ and $\boldsymbol{d_h} = (d_1, \ldots, d_m)$.

The GLS estimator minimizing the GLS distance function $(\boldsymbol{w}-\boldsymbol{d})^T\boldsymbol{\Pi}^{-1}(\boldsymbol{w}-\boldsymbol{d})$ subject to the linear constraints $\boldsymbol{Z}^T\boldsymbol{w}^{\text{ZIE}} = \boldsymbol{T_Z}$ provides the calibrated weights

$$\underset{(n+m)}{\boldsymbol{w}^{\text{ZIE}}} = \boldsymbol{d} + \boldsymbol{\Pi Z}(\boldsymbol{Z}^T\boldsymbol{\Pi Z})^{-1}(\boldsymbol{T_Z} - \boldsymbol{Z}^T\boldsymbol{d}). \tag{4.34}$$

Superscript ZIE indicates Zieschang. Note that $\boldsymbol{w}^{\text{ZIE}} = (w_p^T, w_h^T)^T$ simultaneously delivers person- and household-level weights.

To compare the GLS weights to our alternative weighting approaches, we translate the GLS estimator into the GREG estimator framework. Subsequently, we aim at quantifying the effect caused by the consistency requirements as done for our alternative weighting approaches. The following result proves the equivalence of the GLS estimator introduced by Zieschang (1986, 1990) to a combined GREG estimator based on a combined data set.

Result 7. ***Equivalence of the GLS Estimator to a Combined GREG Estimator***
The weights produced by the GLS estimator introduced by Zieschang (1986, 1990)

$$\boldsymbol{w}^{ZIE} = \boldsymbol{d} + \boldsymbol{\Pi Z}(\boldsymbol{Z}^T\boldsymbol{\Pi Z})^{-1}(\boldsymbol{T_Z} - \boldsymbol{Z}^T\boldsymbol{d})$$

are asymptotically equivalent to the weights produced by the combined GREG estimators

$$\hat{T}_{y_p}^{ZIE} = \hat{T}_{y_p}^{GREG} - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}^{\ T}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}) \tag{4.35}$$

and

$$\hat{T}_{y_h}^{ZIE} = \hat{T}_{y_h}^{GREG} + \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}^{\ T}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}) \tag{4.36}$$

with $\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}$ as person- and household-level GREG estimators for the common totals with $\boldsymbol{x}_i$ and $\boldsymbol{a_g}$ as auxiliaries, respectively. The person-level coefficient is obtained by

$$\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}} = \left(\sum_{i\in s_p} \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}} \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}T} + \sum_{g\in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}T}\right)^{-1} \sum_{i\in s_p} \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}} r_i^{B_x}$$

with residuals $\boldsymbol{r_i^{F_x}} = \boldsymbol{c_i} - \boldsymbol{\hat{F}_x}^T \boldsymbol{x_i}$ *and* $r_i^{B_x} = y_i - \boldsymbol{\hat{B}_x}^T \boldsymbol{x_i}$ *resulting from regressing the common variables and the variable of interest on the auxiliaries. The household-level coefficient is given by*

$$\boldsymbol{\hat{E}_\kappa} = \left(\sum_{i \in s_p} \boldsymbol{r_i^{F_x}} \boldsymbol{r_i^{F_x}}^T + \sum_{g \in s_h} \boldsymbol{r_g^{F_a}} \boldsymbol{r_g^{F_a}}^T \right)^{-1} \sum_{g \in s_h} \boldsymbol{r_g^{F_a}} r_g^{B_a}$$

with household-level residuals $\boldsymbol{r_g^{F_a}} = \boldsymbol{c_g} - \boldsymbol{\hat{F}_a}^T \boldsymbol{a_g}$ *and* $r_g^{B_a} = y_g - \boldsymbol{\hat{B}_a}^T \boldsymbol{a_g}$ *resulting from regressing the common variables and the variable of interest on the auxiliaries.*

Proof. We start by introducing some notation required to rewrite matrices into vectors. Define $s_c = s_p \cup s_h = \{1, \ldots, n, n+1, \ldots, n+m\}$ as an ordered set containing all persons and households indexed by t. For a clear differentiation between the level-specific and combined information, we use Greek letters for the combined information. Suppose

$$\underset{(Q+K+L)\times 1}{\boldsymbol{\zeta_t}} = \begin{cases} (x_{i1}, \ldots, x_{iQ}, 0, \ldots, 0, c_{i1}, \ldots, c_{iL})^T, & \text{for } t \in \{1, \ldots, n\} \\ (0, \ldots, 0, a_{g1}, \ldots, a_{gK}, -c_{g1}, \ldots, -c_{gL})^T, & \text{for } t \in \{n+1, \ldots, n+m\} \end{cases}$$

is a vector containing both the auxiliary and common variables. The combined variable of interest at the person level, $\gamma_{t,p}$, has to be extended by zero if observation t initially belongs to the household sample. It is denoted as

$$\underset{(1\times 1)}{\gamma_{t,p}} = \begin{cases} y_t, & \text{for } t \in \{1, \ldots, n\} \\ 0, & \text{for } t \in \{n+1, \ldots, n+m\}. \end{cases}$$

The combined variable of interest at the household level, $\gamma_{t,h}$ in turn equals zero if observation t initially belongs to the person-level sample. It is given by

$$\underset{(1\times 1)}{\gamma_{t,h}} = \begin{cases} 0, & \text{for } t \in \{1, \ldots, n\} \\ y_t, & \text{for } t \in \{n+1, \ldots, n+m\}. \end{cases}$$

Suppose $\boldsymbol{T_\zeta} = (\boldsymbol{T_x}, \boldsymbol{T_a}, \boldsymbol{0})^T$ and $\boldsymbol{\hat{T}_\zeta^{\text{HT}}} = (\boldsymbol{\hat{T}_x^{\text{HT}}}, \boldsymbol{\hat{T}_a^{\text{HT}}}, \boldsymbol{\hat{T}_\kappa^{\text{HT}}})^T$ are vectors of dimension $(Q+K+L)$ containing the known and estimated totals of the auxiliary and common variables. Note that $\pi_t = \pi_i = \pi_g$.

In a first step, we embed the GLS estimator into the GREG estimation framework. As outlined in Section 2.3.4, a calibration estimator minimizing a chi-squared distance function is asymptotically equivalent to a GREG estimator based on the combined sample s_c and the combined person- and household-level information. If the objective is to estimate a person-level total, the combined GREG estimator is given by

$$\underset{(1\times 1)}{\hat{T}_{\gamma_p}^{\text{ZIE}}} = \underset{(1\times 1)}{\hat{T}_{\gamma_p}^{\text{HT}}} + \underset{1\times(Q+K+L)}{\boldsymbol{\hat{\Psi}_p}^T} \Big(\underset{(Q+K+L)\times 1}{\boldsymbol{T_\zeta}} - \underset{(Q+K+L)\times 1}{\boldsymbol{\hat{T}_\zeta^{\text{HT}}}} \Big) \tag{4.37}$$

with $\hat{\boldsymbol{\Psi}}_{\boldsymbol{p}} = (\sum_{t\in s_c} \pi_t^{-1}\boldsymbol{\zeta_t}\boldsymbol{\zeta_t}^T)^{-1} \sum_{t\in s_c} \pi_t^{-1}\boldsymbol{\zeta_t}\boldsymbol{\gamma_{t,p}}^T$ containing the person-level coefficients. If the objective is to estimate a household-level total, the combined GREG estimator is defined as

$$\underset{(1\times1)}{\hat{T}^{\text{ZIE}}_{\gamma_h}} = \underset{(1\times1)}{\hat{T}^{\text{HT}}_{\gamma_h}} + \underset{1\times(Q+K+L)}{\hat{\boldsymbol{\Psi}}_{\boldsymbol{h}}^{T}} (\underset{(Q+K+L)\times1}{\boldsymbol{T_\zeta}} - \underset{(Q+K+L)\times1}{\hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{\zeta}}}) \tag{4.38}$$

with $\hat{\boldsymbol{\Psi}}_{\boldsymbol{h}} = (\sum_{t\in s_c} \pi_t^{-1}\boldsymbol{\zeta_t}\boldsymbol{\zeta_t}^T)^{-1} \sum_{t\in s_c} \pi_t^{-1}\boldsymbol{\zeta_t}\boldsymbol{\gamma_{t,h}}^T$ containing the household-level coefficients. Therefore, the only difference between $\hat{T}^{\text{ZIE}}_{\gamma_p}$ and $\hat{T}^{\text{ZIE}}_{\gamma_h}$ is given by the combined variable of interest $\boldsymbol{\gamma_p}$ or $\boldsymbol{\gamma_h}$.

In a second step, we are interested in the impact caused by the consistency requirements. Thus, $\boldsymbol{\zeta}_t$ is partitioned into the auxiliary and common variables. Define

$$\underset{(Q+K)\times1}{\boldsymbol{\delta_t}} = \begin{cases} (x_{t1}, \dots, x_{tQ}, 0, \dots, 0)^T, & \text{for } t \in \{1, \dots, n\} \\ (0, \dots, 0, a_{t1}, \dots, a_{tK})^T, & \text{for } t \in \{n+1, \dots, n+m\} \end{cases}$$

as combined auxiliary vector and

$$\underset{(L\times1)}{\boldsymbol{\kappa}_t} = \begin{cases} (c_{tl}, \dots, c_{tL})^T, & \text{for } t \in \{1, \dots, n\} \\ -(\sum\limits_{k\in U_t} c_{kl}, \dots, \sum\limits_{k\in U_t} c_{kL})^T, & \text{for } t \in \{n+1, \dots, n+m\} \end{cases}$$

as vector containing the common variables. It is valid that $\sum_{t\in s_c} \boldsymbol{\delta_t} = \sum_{i\in s_p} \boldsymbol{x_i} + \sum_{g\in s_h} \boldsymbol{a_g}$.

Then, given $\boldsymbol{\delta_t}$ and $\boldsymbol{\kappa_t}$, the person-level combined GREG estimator (4.37) can alternatively be expressed by

$$\underset{(1\times1)}{\hat{T}^{\text{ZIE}}_{\gamma_p}} = \underset{(1\times1)}{\hat{T}^{\text{HT}}_{\gamma_p}} + \underset{1\times(Q+K)}{\hat{\boldsymbol{D}}_{\boldsymbol{\delta}}^{T}} (\underset{(Q+K)\times1}{\boldsymbol{T_\delta}} - \underset{(Q+K)\times1}{\hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{\delta}}}) + \underset{(1\times L)}{\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}^{T}} (\underset{(L\times1)}{\boldsymbol{0}} - \underset{(L\times1)}{\hat{\boldsymbol{T}}^{\mathbf{HT}}_{\boldsymbol{\kappa}}}) \tag{4.39}$$

with $\boldsymbol{T_\delta}$, $\hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{T}}^{\text{HT}}_{\boldsymbol{\kappa}}$ in obvious notation. Coefficients $\hat{\boldsymbol{D}}_{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}$ are simultaneously estimated by

$$\begin{pmatrix} \hat{\boldsymbol{D}}_{\boldsymbol{\delta}} \\ \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}} \end{pmatrix} = \left[\sum_{t\in s_c} \pi_t^{-1} \begin{pmatrix} \boldsymbol{\delta_t} \\ \gamma_t \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta_t} \\ \gamma_t \end{pmatrix}^T \right]^{-1} \sum_{t\in s_c} \pi_t^{-1} \begin{pmatrix} \boldsymbol{\delta_t} \\ \gamma_t \end{pmatrix} \gamma_{t,p}.$$

Analogously to the proceeding in Section 4.2, we decompose $\hat{\boldsymbol{D}}_{\boldsymbol{\delta}}$ using an orthogonal decomposition (cf. Seber, 1977) into

$$\underset{(Q+K)\times1}{\hat{\boldsymbol{D}}_{\boldsymbol{\delta}}} = \underset{(Q+K)\times1}{\hat{\boldsymbol{B}}_{\boldsymbol{\delta}}} - \underset{(Q+K)\times L}{\hat{\boldsymbol{F}}_{\boldsymbol{\delta}}} \underset{(L\times1)}{\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}} \tag{4.40}$$

where $\hat{\boldsymbol{B}}_{\boldsymbol{\delta}}$ results from the model

$$\underset{(1\times1)}{\gamma_t^p} = \underset{1\times(Q+K)}{\hat{\boldsymbol{B}}_{\boldsymbol{\delta}}^{T}} \underset{(Q+K)\times1}{\boldsymbol{\delta_t}} + \underset{(1\times1)}{\boldsymbol{r_t^{B_\delta}}}. \tag{4.41}$$

To better comprehend the effect of regressing the combined auxiliaries on the combined variable of interest, we break down (4.41) to

$$\begin{pmatrix} y_i \\ 0 \end{pmatrix} = \begin{pmatrix} B_{x_1} & \dots & B_{x_Q} & 0 & \dots & 0 \\ 0 & \dots & 0 & B_{a_1} & \dots & B_{a_K} \end{pmatrix} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iQ} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} r_i^{B_x} \\ 0 \end{pmatrix}.$$

Thus, $\hat{\boldsymbol{B}}_\delta$ accounts only for the effects of the auxiliaries $\boldsymbol{x_i}$ on the variables of interest y_i. It does not account for the effects of the common variables.

The product of $\hat{\boldsymbol{F}}_\delta$ and $\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}$ in (4.40) captures the effects of the common variables on the variable of interest neglected by $\hat{\boldsymbol{B}}_\delta$. Coefficient matrix $\hat{\boldsymbol{F}}_\delta$ arises from regressing the combined auxiliaries on the vector of the combined common variable vector

$$\underset{(L\times 1)}{\boldsymbol{\kappa}_t} = \underset{L\times(Q+K)}{\hat{\boldsymbol{F}}_{\boldsymbol{\delta}}{}^T} \underset{(Q+K)\times 1}{\boldsymbol{\delta_t}} + \underset{(L\times 1)}{\boldsymbol{r_t^{F_\delta}}} \tag{4.42}$$

which can be broken down to

$$\begin{pmatrix} c_{t1} \\ \vdots \\ c_{tL} \end{pmatrix} = \begin{pmatrix} F_{x_1}^{c_1} & \dots & F_{x_Q}^{c_1} & F_{a_1}^{c_1} & \dots & F_{a_K}^{c_1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ F_{x_1}^{c_L} & \dots & F_{x_Q}^{c_L} & F_{a_1}^{c_L} & \dots & F_{a_K}^{c_L} \end{pmatrix} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iQ} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} r_t^{F_x^{c_1}} \\ \vdots \\ r_t^{F_x^{c_L}} \end{pmatrix}.$$

Hence, $\hat{\boldsymbol{F}}_\delta$ describes the extent to which the person-level information of $\boldsymbol{x}$ helps to predict the person- and household-level common variables. Inserting the orthogonal decomposition (4.40) into (4.39), we obtain

$$\begin{aligned} \hat{T}_{\gamma_p}^{\text{ZIE}} &= \underbrace{\hat{T}_{\gamma_p}^{\text{HT}} + \hat{\boldsymbol{B}}_{\boldsymbol{\delta}}{}^T (\boldsymbol{T}_\delta - \hat{\boldsymbol{T}}_{\boldsymbol{\delta}}^{\text{HT}})}_{\hat{T}_{\gamma_p}^{\text{GREG}}} - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T \hat{\boldsymbol{F}}_\delta{}^T (\boldsymbol{T}_\delta - \hat{\boldsymbol{T}}_{\boldsymbol{\delta}}^{\text{HT}}) + \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T (\boldsymbol{0} - \hat{\boldsymbol{T}}_{\boldsymbol{\kappa}}^{\text{HT}}) \\ &= \hat{T}_{\gamma_p}^{\text{GREG}} - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T \Big(\hat{\boldsymbol{T}}_{\boldsymbol{\kappa}}^{\text{HT}} + \hat{\boldsymbol{F}}_\delta{}^T (\boldsymbol{T}_\delta - \hat{\boldsymbol{T}}_{\boldsymbol{\delta}}^{\text{HT}}) \Big). \end{aligned}$$

Breaking down the combined vectors and matrices yields

$$\begin{aligned} &= \hat{T}_{\gamma_p}^{\text{GREG}} - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T \left\{ \left(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{HT}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{HT}} \right) + \begin{pmatrix} \hat{\boldsymbol{F}}_{\boldsymbol{x}} \\ \hat{\boldsymbol{F}}_{\boldsymbol{a}} \end{pmatrix} \left[\begin{pmatrix} \boldsymbol{T_x} \\ \boldsymbol{T_a} \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\boldsymbol{HT}} \\ \hat{\boldsymbol{T}}_{\boldsymbol{a}}^{\text{HT}} \end{pmatrix} \right] \right\} \\ &= \hat{T}_{y_p}^{\text{GREG}} - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T (\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}). \end{aligned}$$

According to Remark 3, the partial coefficient $\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}$ can be rewritten as

$$\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}} = \Big(\sum_{t\in s_c} \boldsymbol{r}_t^{\boldsymbol{F}_\delta}\boldsymbol{r}_t^{\boldsymbol{F}_\delta T}\Big)^{-1} \sum_{t\in s_c} \boldsymbol{r}_t^{\boldsymbol{F}_\delta}\boldsymbol{r}_t^{\boldsymbol{B}_\delta}.$$

Finally, given that $\sum_{t\in s_c} \boldsymbol{\delta_t} = \sum_{i\in s_p} \boldsymbol{x_i} + \sum_{g\in s_h} \boldsymbol{a_g}$ and inserting residuals (4.41) and (4.42) into $\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}$, we obtain

$$= \Big(\sum_{i\in s_p} \boldsymbol{r}_i^{\boldsymbol{F}_x}\boldsymbol{r}_i^{\boldsymbol{F}_x T} + \sum_{g\in s_h} \boldsymbol{r}_g^{\boldsymbol{F}_a}\boldsymbol{r}_g^{\boldsymbol{F}_a T}\Big)^{-1} \begin{pmatrix} \sum_{i\in s_p} \boldsymbol{r}_i^{\boldsymbol{F}_x}\boldsymbol{r}_i^{\boldsymbol{B}_x T} \\ 0 \end{pmatrix}.$$

Thus, the person-level estimator (4.35) is proven.

We continue with deducing the GLS estimator at the household-level (4.36). The proceeding is the same as for the person-level estimator. Given $\boldsymbol{\delta_t}$ and $\boldsymbol{\kappa_t}$ and using an orthogonal decomposition, (4.36) can be expressed by

$$\hat{T}^{\text{ZIE}}_{\gamma_h} = \hat{T}^{\text{GREG}}_{y_h} + \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})$$

with

$$\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}} = \Big(\sum_{i\in s_p} \boldsymbol{r}_i^{\boldsymbol{F}_x}\boldsymbol{r}_i^{\boldsymbol{F}_x T} + \sum_{g\in s_h} \boldsymbol{r}_g^{\boldsymbol{F}_a}\boldsymbol{r}_g^{\boldsymbol{F}_a T}\Big)^{-1} \begin{pmatrix} 0 \\ -\sum_{g\in s_h} \boldsymbol{r}_g^{\boldsymbol{F}_a}\boldsymbol{r}_g^{\boldsymbol{B}_a T} \end{pmatrix}.$$

Therefore, Result 7 is proven. □

An essential feature of the GLS estimator is that both the person- and household-level estimators use the same combined auxiliary information $\boldsymbol{\zeta_t}$. The only difference is given by the variable of interest. The impact of ensuring consistency is quantified by the second terms in (4.35) and (4.36) and depends on the difference between the estimated common variable totals and the corresponding coefficients.

The weights of the GLS estimators (4.35) and (4.36) are defined by

$$\begin{aligned} w_i^{\text{ZIE}} &= w_i^{\text{GREG}} - \boldsymbol{r}_i^{\boldsymbol{F}_x}\Big(\sum_{i\in s_p} \boldsymbol{r}_i^{\boldsymbol{F}_x}\boldsymbol{r}_i^{\boldsymbol{F}_x T} + \sum_{g\in s_h} \boldsymbol{r}_g^{\boldsymbol{F}_a}\boldsymbol{r}_g^{\boldsymbol{F}_a T}\Big)^{-1}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}}) \\ w_g^{\text{ZIE}} &= w_g^{\text{GREG}} + \boldsymbol{r}_g^{\boldsymbol{F}_a}\Big(\sum_{i\in s_p} \boldsymbol{r}_i^{\boldsymbol{F}_x}\boldsymbol{r}_i^{\boldsymbol{F}_x T} + \sum_{g\in s_h} \boldsymbol{r}_g^{\boldsymbol{F}_a}\boldsymbol{r}_g^{\boldsymbol{F}_a T}\Big)^{-1}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}}). \end{aligned} \tag{4.43}$$

Given these weights, it can be shown that the GLS estimator is consistent in terms of both the totals of the auxiliary and common variables. We start by verifying the consistency in terms of the known population totals. Given $\sum_{i\in s_p} \boldsymbol{x_i}\boldsymbol{r}_i^{\boldsymbol{F}_x T} = \boldsymbol{0}$ as well as $\sum_{g\in s_h} \boldsymbol{a_g}\boldsymbol{r}_g^{\boldsymbol{F}_a T} = \boldsymbol{0}$, known from the least squares theory (cf. Greene, 2003, Section 6.4 or Wooldridge, 2013, Section 3.2), it is easy to show that the weights (4.43) simultaneously satisfy $\sum_{i\in s_p} w_i^{\text{ZIE}}\boldsymbol{x_i} = \boldsymbol{T_x}$ and

$\sum_{g \in s_h} w_g^{\text{ZIE}} \boldsymbol{a_g} = \boldsymbol{T_a}$. This implies that the sums of the weighted auxiliaries meet the known totals at both levels. We continue with deducing the consistency between the estimated common variable totals. Following Merkouris (2004, p. 1134), it can be shown that

$$\begin{aligned}
\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{ZIE}} &= \sum_{i \in s_p} w_i^{\text{ZIE}} \boldsymbol{c_i} \\
&= \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \sum_{i \in s_p} \boldsymbol{c_i} \boldsymbol{r_i^{F_x}}^T \Big(\sum_{i \in s_p} \boldsymbol{r_i^{F_x}} \boldsymbol{r_i^{F_x}}^T + \sum_{g \in s_h} \boldsymbol{r_g^{F_a}} \boldsymbol{r_g^{F_a}}^T\Big)^{-1} \big(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}\big)
\end{aligned}$$

inserting $\sum_{i \in s_p} \boldsymbol{c_i} \boldsymbol{r_i^{F_x}}^T = \sum_{i \in s_p} \boldsymbol{r_i^{F_x}} \boldsymbol{r_i^{F_x}}^T$ yields

$$\begin{aligned}
&= \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \left[\boldsymbol{1} - \sum_{i \in s_p} \boldsymbol{r_g^{F_a}} \boldsymbol{r_g^{F_a}}^T \Big(\sum_{i \in s_p} \boldsymbol{r_i^{F_x}} \boldsymbol{r_i^{F_x}}^T + \sum_{g \in s_h} \boldsymbol{r_g^{F_a}} \boldsymbol{r_g^{F_a}}^T\Big)^{-1}\right] \big(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}\big) \\
&= \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}} + \sum_{g \in s_h} \boldsymbol{c_g} \boldsymbol{r_g^{F_a}}^T \Big(\sum_{i \in s_p} \boldsymbol{r_i^{F_x}} \boldsymbol{r_i^{F_x}}^T + \sum_{g \in s_h} \boldsymbol{r_g^{F_a}} \boldsymbol{r_g^{F_a}}^T\Big)^{-1} \big(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}\big) \\
&= \sum_{g \in s_h} w_g^{\text{ZIE}} \boldsymbol{c_g} \\
&= \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{ZIE}}.
\end{aligned}$$

Therefore, the GLS approach guarantees consistency between the person- and household-level estimates by the construction of $\boldsymbol{\gamma_t}$, $\boldsymbol{\zeta_t}$ and $\boldsymbol{T_\zeta} = (\hat{\boldsymbol{T}}_{\boldsymbol{x}}^T, \hat{\boldsymbol{T}}_{\boldsymbol{a}}^T, \boldsymbol{0}^T)^T$.

4.3.1.2 Variance Estimation

Zieschang (1990, p. 996) suggested applying balanced repeated replication to estimate the variance for the GLS estimator. However, to be comparable to our alternative weighting approaches, we aim at deriving an analytical expression of the variance. For this purpose, we proceed analogously to Section 4.2 and approximate the nonlinear GLS estimator by linear functions. Due to the analogy with Result 5, we refrain from deriving the formulas in detail and present only the results. The interested reader is referred to Section B.3 in Appendix B for details.

Variance Estimation for Ordinary Variables

The variance estimator of the GLS estimator at the person level (4.35) using the Taylor linearization technique is given by

$$\hat{V}(\hat{T}_{y_p}^{\text{ZIE}}) \doteq \widehat{V}_1 + \widehat{V}_2 + \widehat{V}_3 - 2\widehat{\mathrm{V}}_{12} + 2\widehat{\mathrm{V}}_{13} - 2\widehat{\mathrm{V}}_{23}, \tag{4.44}$$

with

$$\begin{aligned}
\hat{V}_1 &= \hat{V}(\hat{T}_{y_p}^{\text{GREG}}), & \hat{\mathrm{V}}_{12} &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}^T \widehat{\mathrm{Cov}}(\hat{T}_{y_p}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}), \\
\hat{V}_2 &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}) \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}, & \hat{\mathrm{V}}_{13} &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}^T \widehat{\mathrm{Cov}}(\hat{T}_{y_p}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}), \\
\hat{V}_3 &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}, & \hat{\mathrm{V}}_{23} &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}^T \widehat{\mathrm{Cov}}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}.
\end{aligned}$$

At the household level, the variance estimator of the GLS estimator (4.36) using the Taylor linearization technique is obtained from

$$\hat{V}(\hat{T}^{\text{ZIE}}_{y_h}) \doteq \hat{V}_1 + \hat{V}_2 + \hat{V}_3 + 2\hat{\mathrm{V}}_{12} - 2\hat{\mathrm{V}}_{13} - 2\hat{\mathrm{V}}_{23} \tag{4.45}$$

with

$$\begin{aligned}
\hat{V}_1 &= \hat{V}(\hat{T}^{\text{GREG}}_{y_h}), & \hat{\mathrm{V}}_{12} &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T}\,\widehat{\mathrm{Cov}}(\hat{T}^{\text{GREG}}_{y_h}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}}),\\
\hat{V}_2 &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T}\,\hat{V}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}})\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}, & \hat{\mathrm{V}}_{13} &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T}\,\widehat{\mathrm{Cov}}(\hat{T}^{\text{GREG}}_{y_h}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}}),\\
\hat{V}_3 &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T}\,\hat{V}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}, & \widehat{\mathrm{Cov}}_{23} &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T}\,\widehat{\mathrm{Cov}}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}.
\end{aligned}$$

$\widehat{\mathrm{Cov}}$ denotes the estimated covariance. Estimated variances and covariances can be obtained in (2.10) by inserting the appropriate variables.

Variance Estimation for Common Variables

When inserting the common variables as variables of interest into (4.35), the following person-level estimators result

$$\begin{aligned}
\hat{\boldsymbol{T}}^{\text{ZIE}}_{\boldsymbol{c_p}} &= \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})\\
&= (1 - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^{T})\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} + \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^{T}\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}}.
\end{aligned}$$

Accordingly, the person-level estimator for the common variables can be written as a composite estimator with the single estimates weighted by $\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}$. In contrast to $\hat{\boldsymbol{D}}_{\boldsymbol{c}}$ from our proposed weighting approaches, $\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}$ is not given by a diagonal matrix. The reason is that $\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}$ arises from a model with a vector containing the person-level common variables and zeros on the left-hand side of the model and a matrix containing the person- and household-level common variable information on the right-hand side. Therefore, it is evident that the person-level information is not completely explained by the combined person- and household-level information.

The corresponding variance estimator is given by

$$\begin{aligned}
V(\hat{\boldsymbol{T}}^{\text{ZIE}}_{\boldsymbol{c_p}}) = {} & (1 - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^{T})V(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}})(1 - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}) + \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^{T}V(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}\\
& + 2(1 - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^{T})\mathrm{Cov}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}.
\end{aligned}$$

Inserting the common variables into the household-level estimator (4.36), we obtain the following composite estimator

$$\begin{aligned}
\hat{\boldsymbol{T}}^{\text{ZIE}}_{\boldsymbol{c_h}} &= \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}} + \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})\\
&= (1 - \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T})\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}} + \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T}\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}}.
\end{aligned}$$

The corresponding variance estimator is obtained from

$$\begin{aligned}
V(\hat{\boldsymbol{T}}^{\text{ZIE}}_{\boldsymbol{c_h}}) = {} & (1 - \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T})V(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})(1 - \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}) + \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T}V(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}})\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}\\
& + 2(1 - \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T})\mathrm{Cov}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}.
\end{aligned}$$

Compared with the variance formulas of our proposed alternative estimators, introduced in Section 4.2, $V(\hat{\boldsymbol{T}}^{\text{ZIE}}_{\boldsymbol{c_p}})$ and $V(\hat{\boldsymbol{T}}^{\text{ZIE}}_{\boldsymbol{c_h}})$ are computationally more demanding, because the variance estimators at both levels and their covariances are required.

4.3.2 GLS Estimator According to Merkouris (2004)

Merkouris (2004) modified the GLS estimator to account for the effective sample sizes of the independent multiple surveys. We only briefly consider the modified estimator, because the conceptual differences from the original GLS estimator proposed by Zieschang (1986, 1990) are minor. Suppose

$$\pi_t = \begin{cases} \text{deff}_{s_p}/\pi_t n & \text{for } \; t \in \{1, \dots, n\} \\ \text{deff}_{s_h}/\pi_t m & \text{for } \; t \in \{n+1, \dots, n+m\}. \end{cases}$$

as the combined inclusion probability with deff_s as design effect of sample s. The design effect describes the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements (cf. Kish, 1965, p. 258). Inserting π_t into (4.35) and (4.36) yields the following modified estimators

$$\hat{T}_{y_p}^{\text{MER}} = \hat{T}_{y_p}^{\text{GREG}} - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}^{\text{MER}^T} (\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \tag{4.46}$$

and

$$\hat{T}_{y_h}^{\text{MER}} = \hat{T}_{y_h}^{\text{GREG}} + \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}^{\text{MER}^T} (\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \tag{4.47}$$

with $\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}$ and $\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}$ as person- and household-level GREG estimators for the common totals with $\boldsymbol{x_i}$ and $\boldsymbol{a_g}$ as auxiliaries, respectively. Superscript MER refers to Merkouris. The coefficients are obtained from

$$\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}^{\text{MER}} = \left((1-q) \sum_{i \in s_p} \boldsymbol{r_i^{F_x}} \boldsymbol{r_i^{F_x}}^T + q \sum_{g \in s_h} \boldsymbol{r_g^{F_a}} \boldsymbol{r_g^{F_a}}^T \right)^{-1} (1-q) \sum_{i \in s_p} \boldsymbol{r_i^{F_x}} r_g^{B_x} \tag{4.48}$$

and

$$\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}^{\text{MER}} = \left((1-q) \sum_{i \in s_p} \boldsymbol{r_i^{F_x}} \boldsymbol{r_i^{F_x}}^T + q \sum_{g \in s_h} \boldsymbol{r_g^{F_a}} \boldsymbol{r_g^{F_a}}^T \right)^{-1} q \sum_{g \in s_h} \boldsymbol{r_g^{F_a}} r_g^{B_a}, \tag{4.49}$$

where the weighting factor $q = \dfrac{n/\text{deff}_{s_p}}{n/\text{deff}_{s_p} + m/\text{deff}_{s_h}}$ is proportional to the effective sample size. Therefore, the modified estimators introduced by Merkouris (2004) differ from the original GLS estimators (4.35) and (4.36) only with respect to weighting factor q. The variance estimators arises considering the weighting factor q in (4.44) and (4.45), as done in (4.48) and (4.49).

4.4 Comparison of Our Alternative Weighting Approaches and the GLS Estimator

This section aims to compare our alternative weighting approaches to the GLS estimator. All estimators under consideration ensure consistency between person- and household-level estimates. Table 4.2 summarizes the estimators under consideration, given in a similar expression to make the differences more accessible. It becomes evident that the differences arise from the signs of the adjustment terms, the formulas of the coefficients, and the estimated common variable totals.

Moreover, the two approaches differ conceptually on how consistency is ensured. Our weighting approaches use the same estimated common variable totals in both the person- and household-level estimator. We suggest a person-level estimator as estimator for the unknown common variable totals, since in household surveys it is more prevalent that the common variables are initial person characteristics that are assigned in aggregated form to the household-level data set. In contrast, the GLS estimator enforces consistent person- and household-level estimates more indirectly through the construction of the combined variable of interest $\boldsymbol{\gamma}_t$ and the auxiliary information $\boldsymbol{\zeta}_t$ as well as through the known total vector $\boldsymbol{T}_\zeta = (\hat{\boldsymbol{T}}_{\boldsymbol{x}}^T, \hat{\boldsymbol{T}}_{\boldsymbol{a}}^T, \boldsymbol{0}^T)^T$. The final estimates of the unknown common variable totals are determined by a weighted average of the single person- and household-level estimates. Therefore, the same common variable information is used twice, once in its initial form at the person level and once in aggregated form at the household level. However, it is questionable to what extent the aggregated household-level information, supplementary to the person-level information, helps to predict the common variable totals.

Furthermore, the variance estimators of the common variables differ. In our alternative weighting approaches, the variance of the common variables totals depends solely on the person-level variance estimator. In contrast, in the GLS approach the variance estimators of the common variables are more elaborate, since the variance estimators both the person and household level and their covariances are required.

In addition, the number of calculation steps differs. The GLS estimator is a one-step procedure. Our proposed weighting approaches consist of two calculation steps. In a first step, the unknown common variable totals are estimated. In a second step, based on the estimated common variable totals from the first step, the final estimators are determined.

Finally, when comparing the coefficients of the adjustment terms accounting for the impact of consistency, it becomes evident that the combined coefficients $\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}$ and $\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}$ simultaneously use person- and household-level information by the term $(\sum_{i \in s_p} \boldsymbol{r}_i^{\boldsymbol{F_x}} \boldsymbol{r}_i^{\boldsymbol{F_x}T} + \sum_{g \in s_h} \boldsymbol{r}_g^{\boldsymbol{F_a}} \boldsymbol{r}_g^{\boldsymbol{F_a}T})$. However, as mentioned in the previous paragraph, in the context of household surveys, it seems questionable to what extent the household-level auxiliary information helps to predict the person-level variables.

Table 4.2: Summary of the proposed and benchmark estimators II

Person-level	Household-level
First proposed weighting approach (WA1)	
$\hat{T}^{\text{WA1}}_{y_p} = \hat{T}^{\text{GREG}}_{y_p}$	$\hat{T}^{\text{WA1}}_{y_h} = \hat{T}^{\text{GREG}}_{y_h} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})$ with $\hat{\boldsymbol{E}}_{\boldsymbol{c}} = (\sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}}{}^{T})^{-1} \sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} \boldsymbol{r}^{\boldsymbol{B_a}}_{\boldsymbol{g}}$
Second proposed weighting approach (WA2)	
$\hat{T}^{\text{WA2}}_{y_p} = \hat{T}^{\text{GREG}}_{y_p} + \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c^*_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}})$ with $\hat{\boldsymbol{D}}_{\boldsymbol{c}} = (\sum_{i\in s_p} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}}{}^{T})^{-1} \sum_{i\in s_p} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}} r^{B_x}_{i}$	$\hat{T}^{\text{WA2}}_{y_h} = \hat{T}^{\text{GREG}}_{y_h} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c^*_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})$ with $\hat{\boldsymbol{E}}_{\boldsymbol{c}} = (\sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}}{}^{T})^{-1} \sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} \boldsymbol{r}^{\boldsymbol{B_a}}_{\boldsymbol{g}}$
GLS estimator by Zieschang (1986, 1990) (ZIE) as benchmark	
$\hat{T}^{\text{ZIE}}_{y_p} = \hat{T}^{GREG}_{y_p} - \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^{T}(\hat{\boldsymbol{T}}^{\boldsymbol{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\boldsymbol{GREG}}_{\boldsymbol{c_h}})$ with $\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}} = (\sum_{i\in s_p} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}}{}^{T} + \sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}}{}^{T})^{-1} \sum_{i\in s_p} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}} r^{B_x}_{i}$	$\hat{T}^{\text{ZIE}}_{y_h} = \hat{T}^{GREG}_{y_h} + \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^{T}(\hat{\boldsymbol{T}}^{\boldsymbol{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\boldsymbol{GREG}}_{\boldsymbol{c_h}})$ with $\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}} = (\sum_{i\in s_p} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}}{}^{T} + \sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}}{}^{T})^{-1} \sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} r^{B_a}_{g}$
GLS estimator by Merkouris (2004) (MER) as benchmark	
$\hat{T}^{\text{MER}}_{y_p} = \hat{T}^{\text{GREG}}_{y_p} - \hat{\boldsymbol{D}}^{M}_{\boldsymbol{\kappa}}{}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})$ with $\hat{\boldsymbol{D}}^{\text{MER}}_{\boldsymbol{\kappa}} = \left((1-q)\sum_{i\in s_p} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}}{}^{T} + q\sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}}{}^{T}\right)^{-1}$ $(1-q)\sum_{i\in s_p} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}} r^{B_x}_{g}$	$\hat{T}^{\text{MER}}_{y_h} = \hat{T}^{\text{GREG}}_{y_h} + \hat{\boldsymbol{E}}^{M}_{\boldsymbol{\kappa}}{}^{T}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})$ with $\hat{\boldsymbol{E}}^{\text{MER}}_{\boldsymbol{\kappa}} = \left((1-q)\sum_{i\in s_p} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}} \boldsymbol{r}^{\boldsymbol{F_x}}_{\boldsymbol{i}}{}^{T} + q\sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}}{}^{T}\right)^{-1}$ $q\sum_{g\in s_h} \boldsymbol{r}^{\boldsymbol{F_a}}_{\boldsymbol{g}} r^{B_a}_{g}$

4.5 Simulation Study

The following MC simulation study compares the performance of our alternative weighting approaches, the GLS estimators as benchmark estimators and integrated weighting as current practice in official offices (as presented in detail in Chapter 3). It should be noted that all estimators under consideration (Table 4.3) ensure consistency between person- and household-level estimates.

Table 4.3: Estimators under consideration

Estimator	Description
	Integrated weighting
INT1	Integrated GREG estimator according to Lemaître and Dufour (1987) determined by Definition 3 with $v_i = 1$
INT2	Integrated GREG estimator according to Nieuwenbroek (1993) determined by Definition 3 with $v_i = N_g^{-1}$
	Alternative weighting approaches
WA1	First alternative weighting approach determined by (4.1) and (4.4)
WA2	First alternative weighting approach determined by (4.19) and (4.25)
	Benchmark estimators
ZIE	GLS estimator proposed by Zieschang (1986, 1990) determined by (4.35) and (4.36)
MER	GLS estimator suggested by Merkouris (2004) defined in (4.46) and (4.47)

The simulation study is based on the same simulation setup as used in the previous chapters (see Section 3.4.1 for details). We draw 1000 MC samples of $m = 1500$ and $m = 200$ households by means of simple random sampling. The auxiliaries consist of the same auxiliary variables as presented in Table 3.6. In the integrated GREG estimators, we also include the additional auxiliary variable N_g^{-1} required to ensure the integrated property (see Section 3.1.1 for a definition). For a fair comparison, we also incorporate the household size as further auxiliary variable into the alternative weighting approaches and into the GLS estimators. The integrated weights are computed at the person level and are then assigned one-to-one to the corresponding household. Following, it is implicitly assumed that the household characteristics are explained by the same auxiliary variables as the person characteristics but in aggregated form. Therefore, in order to be comparable with integrated weighting, we use the same auxiliary variables in our alternative weighting approaches and in the GLS estimators. However, it is important to note that our alternative weighting approaches allow us to utilize different auxiliaries at the person and household level. The variables of interest at the household level are presented in Table 4.4. As common variables, emerging in both the person- and the household-level data set, we choose inc and soc. RB, MSE and rsRB$_r$ introduced in Section 3.4.1 serve as quality measures. The simulation study consists of three parts: Section 4.5.1 investigates the distribution of the weights obtained from the methods under consideration. Sections 4.5.2 and 4.5.3 present the results on point and variance estimates.

Table 4.4: Variables of interest at the household level

Variable	Description
inc	Personal income
soc	Social income
cap_inc	Capital income (interest, dividends, profit from capital investments in unincorporated business)
taxes	Regular taxes on wealth

4.5.1 Results on Weights

The weights of the competing methods divided by the design weight for different sample sizes and for all 1000 MC samples are plotted in Figure 4.1. It becomes apparent that the person weights of our alternative weighting approaches, WA1 and WA2, and of the GLS estimators, ZIE and MER, have a considerably smaller range than the weights of INT1 and INT2. Actually, for $m = 200$, INT1 and INT2 produce a considerable number of negative weights. Since the integrated person weights are assigned one-to-one to the household level, the ranges of the weights of INT1 and INT2 are equal between both levels. However, the interquartile ranges differs. For $m = 200$ the household weights of WA1 and WA2 also vary less than the household weights for INT1. At the household level, negative weights emerge in almost all approaches. However, a considerable body of literature exists on avoiding negative weights, as presented in Section 2.3.5. Weight distributions of WA1, ZIE, and MER are quite similar. Figure 4.2 depicts that, in contrast to INT1 and INT2, the ranges of the person weights of WA1, WA2, ZIE and MER are independent of the household size.

4.5.2 Results on Point Estimates

The empirical biases in Tables B.5 and B.5 in Appendix B confirm that all estimators under consideration are asymptotically unbiased. Table 4.5 summarizes the ratios of the MSEs (see Section 3.4.1 for a definition) of integrated weighting relative to our alternative weighting approaches and relative to the GLS estimators. It becomes evident that all numbers in the table

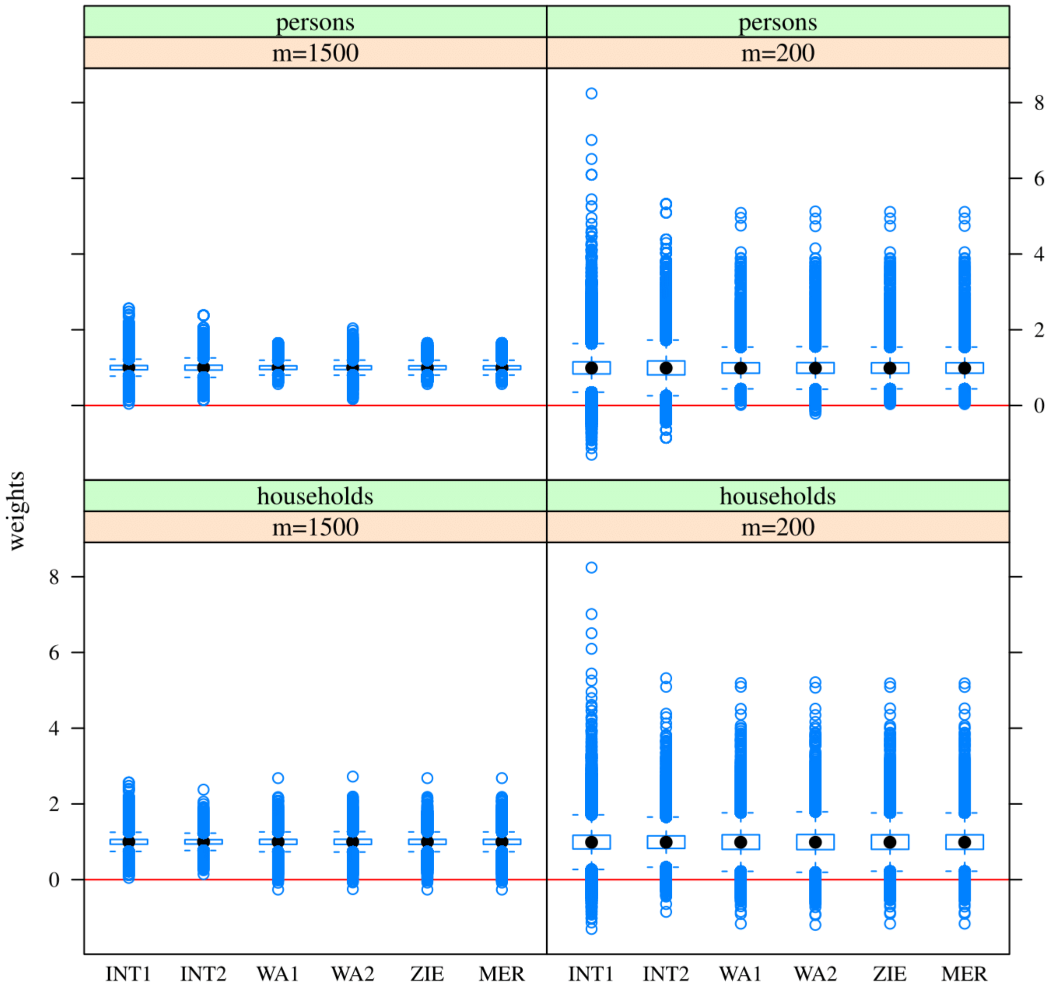

Figure 4.1: Boxplots of the person- and household-level weights

are greater than or at least 1, except from one number for bene_age3 and $m = 200$. Therefore, WA1, WA2, ZIE, and MER perform at least as well as integrated weighting. The greatest gains in precision are realized for WA2. Here, all variables benefit from inserting improved estimates for the common variable totals estimated by a specialized auxiliary variable set (the specialized variables are given B.4 in Appendix B). Actually, the gains in precision for the common variable inc ranges up to 73%. Even if the common variables are included as additional auxiliaries, no considerable precision gains are realized for WA1, ZIE and MER with respect to the common variables. Our results do not confirm the observation of Merkouris (2004, p. 1131) that MER improves the precision compared to ZIE.

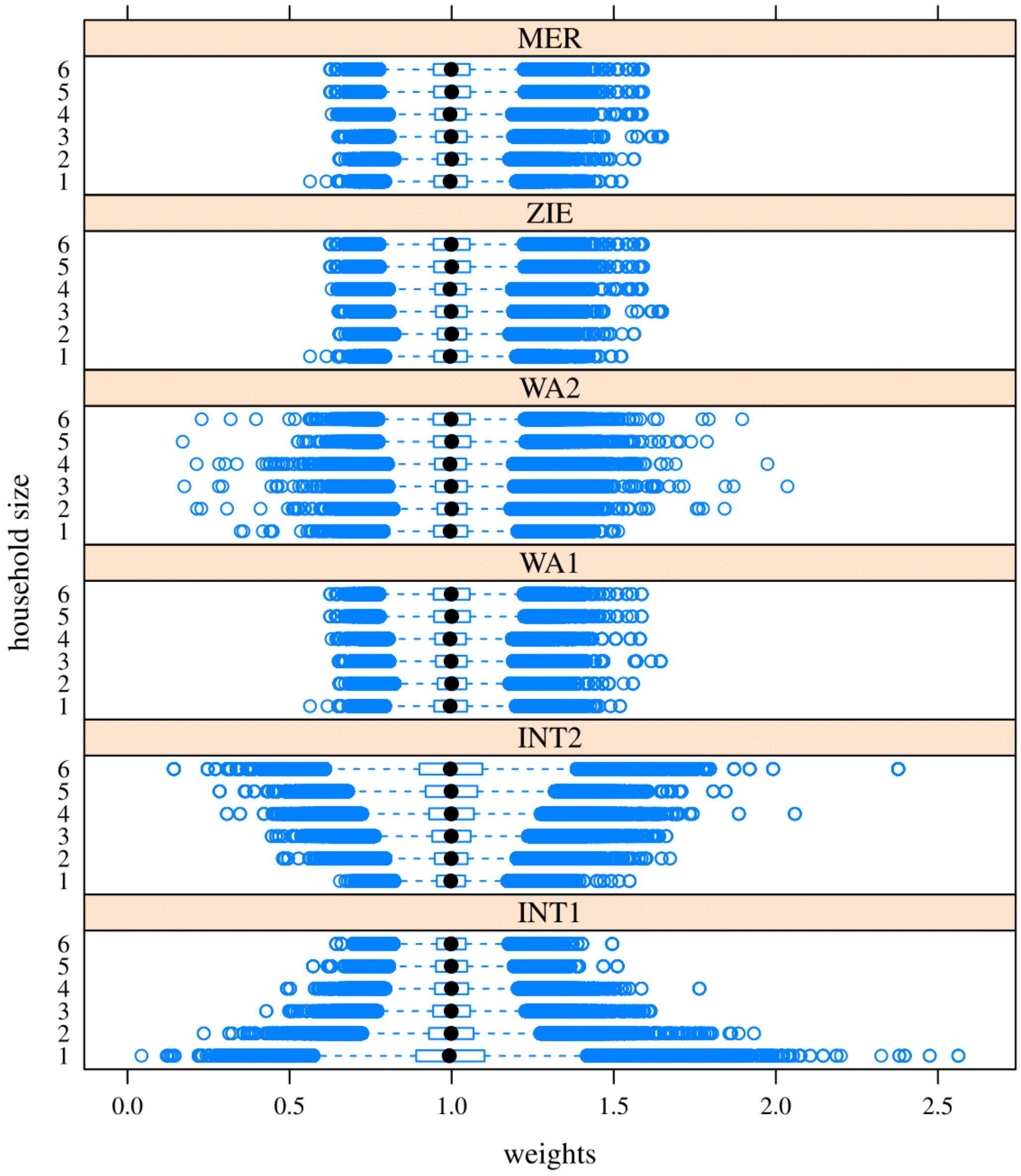

Figure 4.2: Boxplots for person weights by household size for $m = 1500$

Table 4.5: Relative efficiency of the MSE of point estimates at the person level

	m=1500								m=200							
	$\frac{\text{INT1}}{\text{WA1}}$	$\frac{\text{INT2}}{\text{WA1}}$	$\frac{\text{INT1}}{\text{WA2}}$	$\frac{\text{INT2}}{\text{WA2}}$	$\frac{\text{INT1}}{\text{ZIE}}$	$\frac{\text{INT2}}{\text{ZIE}}$	$\frac{\text{INT1}}{\text{MER}}$	$\frac{\text{INT2}}{\text{MER}}$	$\frac{\text{INT1}}{\text{WA1}}$	$\frac{\text{INT2}}{\text{WA1}}$	$\frac{\text{INT1}}{\text{WA2}}$	$\frac{\text{INT2}}{\text{WA2}}$	$\frac{\text{INT1}}{\text{ZIE}}$	$\frac{\text{INT2}}{\text{ZIE}}$	$\frac{\text{INT1}}{\text{MER}}$	$\frac{\text{INT2}}{\text{MER}}$
`inc`	1.00	1.00	1.70	1.70	1.00	1.00	1.00	1.00	1.02	1.02	1.73	1.73	1.01	1.01	1.01	1.01
`soc`	1.01	1.01	1.22	1.22	1.00	1.00	1.00	1.00	1.01	1.02	1.23	1.24	1.01	1.02	1.01	1.02
`sel`	1.00	1.00	1.04	1.04	1.00	1.00	1.00	1.00	1.00	1.01	1.04	1.04	1.00	1.01	1.00	1.01
`act1`	1.00	1.00	1.30	1.30	1.00	1.00	1.00	1.00	1.01	1.01	1.32	1.32	1.01	1.01	1.01	1.01
`act2`	1.00	1.00	1.05	1.05	1.00	1.00	1.00	1.00	1.01	1.02	1.07	1.08	1.01	1.02	1.01	1.02
`act3`	1.00	1.00	1.15	1.15	1.00	1.00	1.00	1.00	1.01	1.01	1.17	1.18	1.01	1.01	1.01	1.01
`bene_age1`	1.00	1.01	1.03	1.03	1.00	1.01	1.00	1.01	1.01	1.00	1.05	1.04	1.02	1.01	1.02	1.01
`bene_age2`	1.00	1.00	1.10	1.10	1.00	1.00	1.00	1.00	1.01	1.02	1.15	1.17	1.01	1.02	1.01	1.02
`bene_age3`	1.00	1.00	1.07	1.07	1.00	1.00	1.00	1.00	0.99	1.00	1.11	1.12	1.00	1.00	1.00	1.00
`bene_age4`	1.00	1.00	1.02	1.02	1.00	1.00	1.00	1.00	1.01	1.02	1.05	1.06	1.01	1.02	1.01	1.02
`inc_hs1`	1.13	1.08	1.17	1.12	1.13	1.08	1.13	1.08	1.24	1.10	1.26	1.11	1.24	1.09	1.24	1.09
`inc_hs2`	1.30	1.28	1.38	1.35	1.30	1.28	1.30	1.28	1.33	1.26	1.34	1.27	1.33	1.26	1.33	1.26
`inc_hs3`	1.38	1.36	1.47	1.46	1.37	1.36	1.37	1.36	1.45	1.44	1.51	1.50	1.45	1.44	1.45	1.44
`inc_hs4`	1.49	1.49	1.64	1.64	1.49	1.49	1.49	1.49	1.45	1.48	1.56	1.60	1.45	1.48	1.45	1.48
`inc_hs5`	1.10	1.11	1.11	1.12	1.10	1.11	1.10	1.11	1.07	1.11	1.11	1.16	1.08	1.12	1.08	1.12
`inc_hs6`	1.11	1.07	1.13	1.09	1.11	1.06	1.11	1.06	1.12	1.15	1.14	1.16	1.12	1.14	1.12	1.14

Table 4.6: Relative efficiency of the MSE of point estimates at the household level

	m=1500								m=200							
	$\frac{\text{INT1}}{\text{WA1}}$	$\frac{\text{INT2}}{\text{WA1}}$	$\frac{\text{INT1}}{\text{WA2}}$	$\frac{\text{INT2}}{\text{WA2}}$	$\frac{\text{INT1}}{\text{ZIE}}$	$\frac{\text{INT2}}{\text{ZIE}}$	$\frac{\text{INT1}}{\text{MER}}$	$\frac{\text{INT2}}{\text{MER}}$	$\frac{\text{INT1}}{\text{WA1}}$	$\frac{\text{INT2}}{\text{WA1}}$	$\frac{\text{INT1}}{\text{WA2}}$	$\frac{\text{INT2}}{\text{WA2}}$	$\frac{\text{INT1}}{\text{ZIE}}$	$\frac{\text{INT2}}{\text{ZIE}}$	$\frac{\text{INT1}}{\text{MER}}$	$\frac{\text{INT2}}{\text{MER}}$
`inc`	1.00	1.00	1.70	1.70	1.00	1.00	1.00	1.00	1.02	1.02	1.73	1.73	1.01	1.01	1.01	1.01
`soc`	1.01	1.01	1.22	1.22	1.00	1.00	1.00	1.00	1.01	1.02	1.23	1.24	1.01	1.02	1.01	1.02
`gross_inc`	1.00	1.00	1.65	1.65	1.00	1.00	1.00	1.00	1.02	1.02	1.69	1.68	1.02	1.01	1.02	1.01
`cap_inc`	1.00	1.01	1.00	1.01	1.00	1.01	1.00	1.01	0.96	0.98	0.96	0.98	0.96	0.98	0.96	0.98
`taxes`	1.00	1.01	1.01	1.01	1.00	1.01	1.00	1.01	1.12	1.11	1.11	1.11	1.12	1.12	1.12	1.12

4.5.3 Results on Variance Estimates

Figure 4.3 plots the rsRB$_r$ for $r = 1, \ldots, 1000$ of the estimated variances at the person level. The two lower rows present the larger sample size with $m = 1500$, the two upper rows show the smaller sample size with $m = 200$. The RB of the variance estimates is indicated in green. For $m = 1500$, the variance estimators of INT1, INT2, and WA1 perform very similar. WA2 also delivers accurate estimates although the variance estimator is more complex than for WA1, INT1, and INT2 since it contains six variance components instead of one, as derived in Section 4.2.2.2. Actually, for bene_age2 and bene_age3, WA2 outperforms all other methods under consideration. Only for inc and all variables related to it, do the distributions of the variance estimates have a wider range. The reason for this is that due to the specialized estimates of the unknown common variable totals more randomness and variation is introduced into the variance estimator. For $m = 200$, the variance estimates of WA2 outperform all estimators under consideration for bene_age1, bene_age2, bene_age3 and bene_age4.

The variance estimators of the benchmark methods ZIE and MER underestimate the empirical variance of the point estimates, particularly for the common variables. This is likely because the Taylor linearization does not capture the additional randomness introduced by the combined estimate of the common variables. As a remedy, Zieschang (1990, p. 996) and Merkouris (2004, p. 1137) recommended resampling methods. However, since ZIE and MER are only benchmark methods, the variance estimation for our proposed weighting approaches works very accurately, and resampling methods are out of the scope of the thesis, we refrain from computing resampling variances.

Larger differences between the estimators under consideration can be found at the household level. Figure 4.4 shows that also at the household level the variance estimation of WA1 works accurately. As at the person level, the boxes of WA2 are wider for the common variables and for variables related to them, as compared to WA1, INT1, and INT2. For taxes, the boxes of WA1 and WA2 are nearly identical. However, INT1 and INT2 produce various outliers and underestimates the empirical variance of taxes, which is characterized by a very skewed distribution with several zeros. In contrast, our proposed weighting approaches WA1 and WA2, as well the GLS estimators ZIE and MER, seem to be robust against outliers. Moreover, also at the household level, the Taylor linearization for ZIE and MER does not produce reliable variance estimates for the common variables.

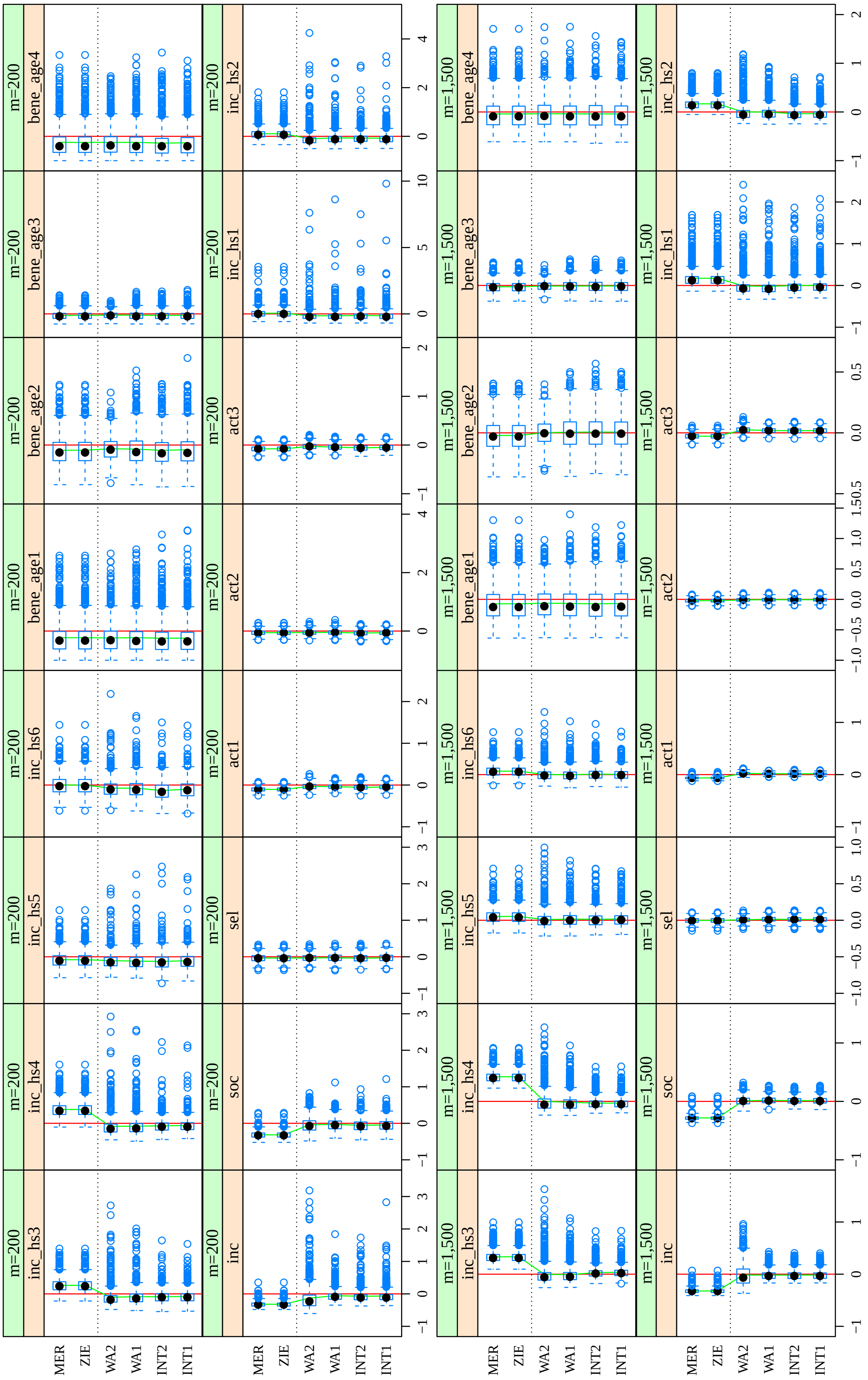

Figure 4.3: Relative Bias and replicate specific-relative bias of the estimated variances at the person level

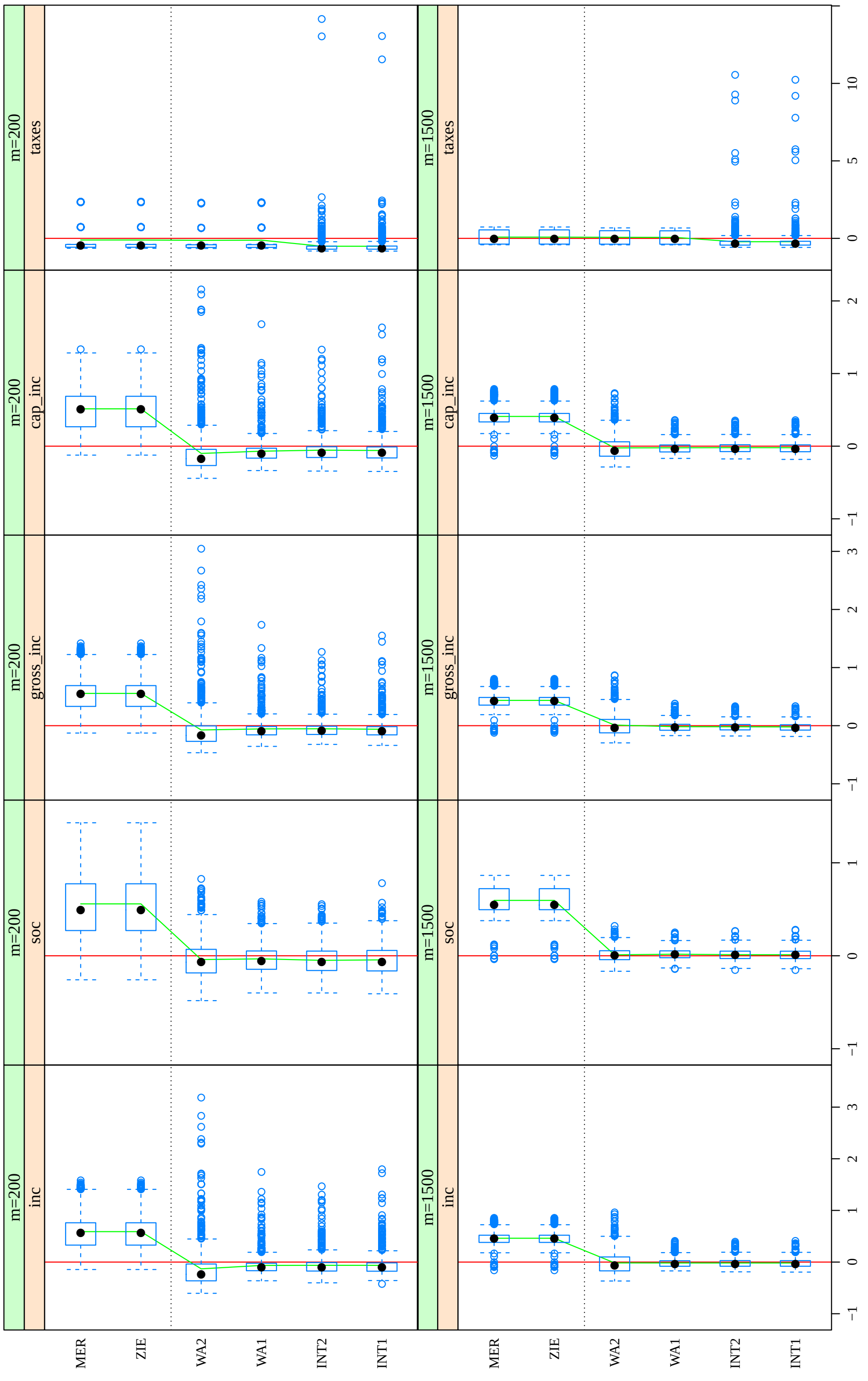

Figure 4.4: Relative Bias and replicate specific-relative bias of the estimated variances at the household level

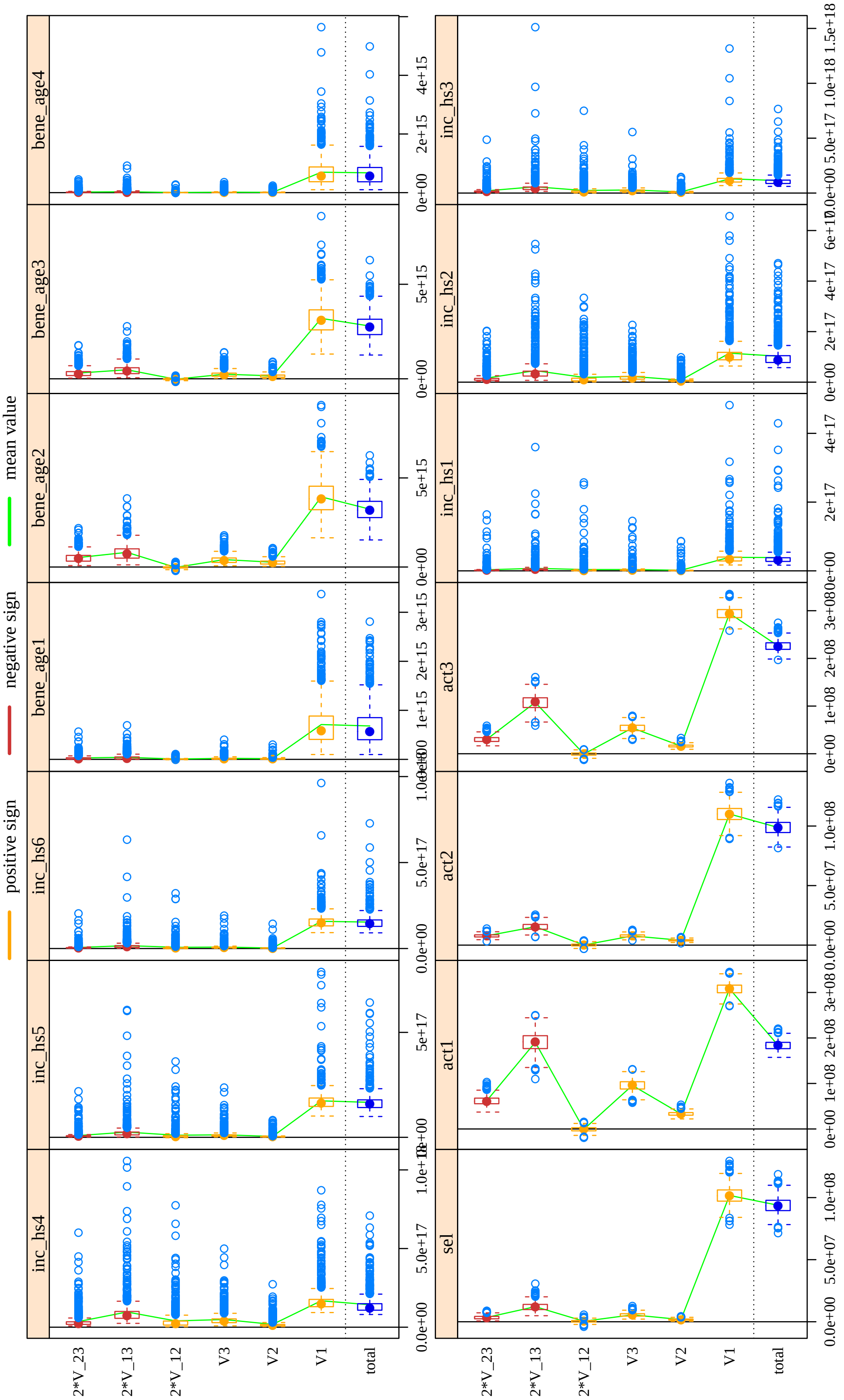

Figure 4.5: Variance components for second weighting approach at the person level

Regarding our alternative weighting approaches, we are interested in the relation between the variance components. Figure 4.5 shows the total variance estimates and the variance components for WA2 at the person level (see (4.19) and (4.28) for the formulas) for all variables of interest except of the common variables. The latter variance estimator comprise only one single variance component (see (4.30) and (4.31)). Variance components with a negative sign are highlighted in red, variance components with a positive sign are highlighted in orange. It can be seen that variance component V_1, describing the variance estimates of the variable of interest, exceeds the total variance estimates of WA2. In other words, the incorporation of common variables as additional auxiliaries decreases the total variance. Variance component V_3 considering the common variable totals estimated by the specialized auxiliary variable set $\boldsymbol{z}$ exceeds V_2, considering the variances estimates of the common variable totals estimated by the auxiliary variable set $\boldsymbol{x}$. This result is not surprising, because the number of variables in $\boldsymbol{z}$ is higher than in $\boldsymbol{x}$ in our simulation setup. Moreover, the covariance terms V_{12} and V_{23} are very small compared to V_{13}. The reason for this is that the latter term concerns the covariance of estimates based on different auxiliary variable sets.

An interesting question is whether the estimation of six instead of one variance components in the variance estimator for our alternative weighting approaches affects their convergence behavior. Figure 4.6 depicts the convergence plots of the rsRB$_r$ for $r = 1, \ldots, 1000$ of the estimated variances for the common variable `inc` at the household level. At this level, the variance estimators of both WA1 and WA2 comprise six variance components. It becomes evident that WA1 achieves nearly the same convergence behavior as the variance estimates of INT1 and INT2. The difference between the variance estimates and the empirical variance of the estimator of WA2 exceeds the difference for the other estimators up to $R < 200$. For $R \geq 200$, the convergence speed adjusts for the competitive variance estimators.

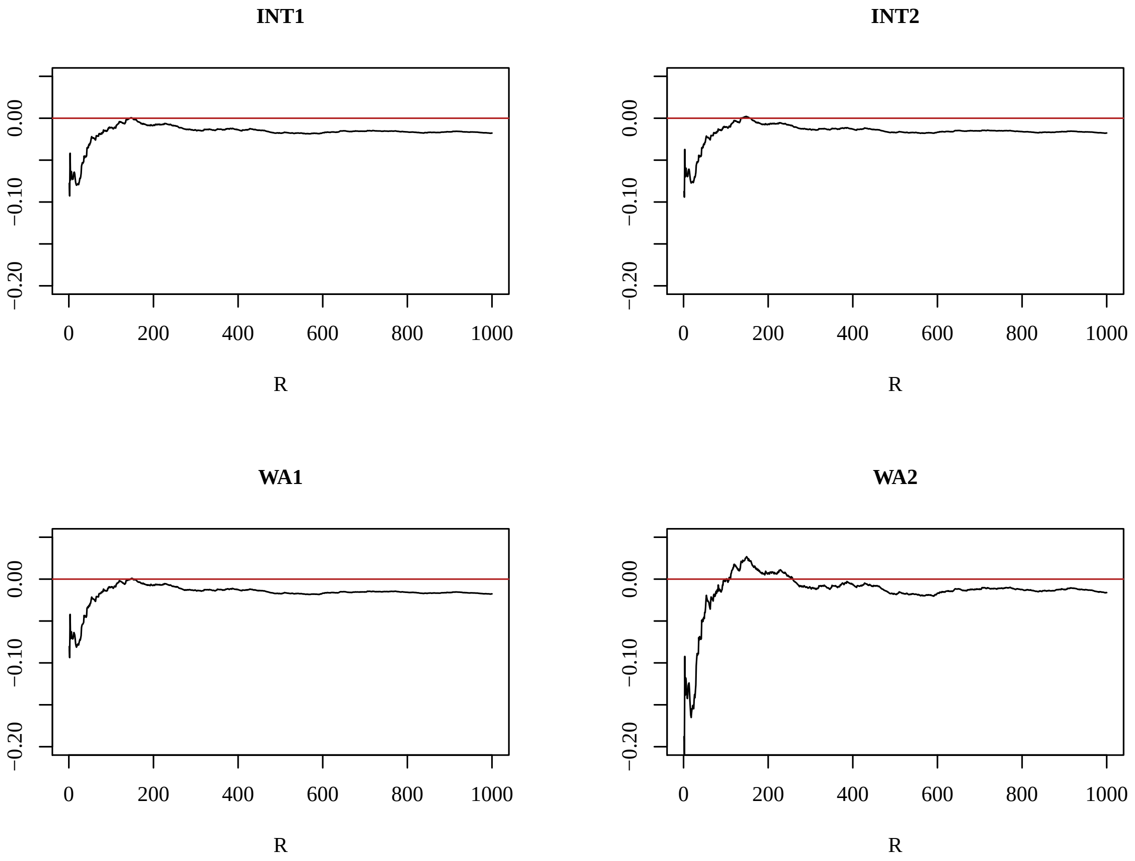

Figure 4.6: Convergence plot of the relative bias of the estimated variances at the household level for `inc` and $m = 1500$

4.6 Summary and Conclusion

In this chapter, we proposed two weighting approaches as alternatives to integrated weighting. These alternative weighting approaches are capable of both ensuring consistent person- and household-level estimates and allowing for different weights for the persons within the same household. The advantages of the alternative weighting approaches compared to integrated weighting are manifold. First, consistency is ensured more directly and only for the relevant variables, instead of indirectly by aggregating the individual information per household. Second, using the original auxiliary information allows divergent weights for the persons within the same household. Therefore, the heterogeneity in a household, if it exists, is captured, and individual patterns are retained. Thirdly, at the person and household level, different models can be implemented, which ensures more flexibility in variable selection and prevents problems induced by ecological fallacy. Finally, no additional auxiliary variable is required to enforce the integrated property.

To ensure consistent person- and household-level estimates, we included the variables common to both the person- and household-level data sets in our weighting approaches as additional auxiliary variables. For this purpose, we extended the method of Renssen and Nieuwenbroek (1997) of combining information of independent multiple surveys. However, there are considerable differences between multiple surveys and household surveys in terms of the definition of common variables, the dependence of the surveys and differing target populations. The difference between our first and second weighting approach is given by the implementation effort and the quality of the estimated totals of the common variables. Whereas the first approach is easier to implement, since only the household-level estimator has to be extended by the common variables, the second weighting approach offers the best available estimate for the unknown common variable totals. The simulation results on point and variance estimates confirm the trade-off between the implementation expense and the quality of the final estimates in the choice between our two weighting approaches. The variance estimators of our proposed weighting approaches account for the additional source of randomness induced by the estimated totals of the common variables. Therefore, we derived the variance estimators for each proposed estimator via Taylor linearization considering the additional randomness.

As a benchmark estimator for the alternative weighting approaches, we adopt the GLS estimator introduced by Zieschang (1986, 1990), which combines information from independent multiple surveys. In this approach consistency between person- and household-level estimates is indirectly ensured by the construction of the pooled auxiliary variables and the linear constraints. Moreover, the estimates for the unknown common variable totals are produced by a weighted average of the separate estimates obtained from each of the surveys. Therefore, the same common variable information is used twice, once in its original person-level form and once in aggregated form at the household level. In contrast, the unknown common variable totals in our weighting approaches are based on the person level in order to account that it is more prevalent that the common variables are initial person-level characteristics.

Our simulation study strongly supports the superiority of our alternative weighting approaches compared to integrated weighting and the GLS estimators. In particular, our second proposed weighting approach yields the most precise point estimates. The precision gains depend on the strength of the relation between the common variables and variables of interest. Our first weighting approach and the GLS estimators perform very similar. As a result, with the proposed alternative weighting approaches we contradict the wide-spread perception in the literature that equal weights are required to ensure consistent estimates.

The advantages of our proposed alternative weighting approaches become more obvious when the weights are adjusted for nonresponse. In general, methods to prevent nonresponse bias proceed at the person level. As a result, the adjusted person weights are no longer necessarily equal within a household. In order to maintain consistency, Eurostat (cf. European Commission, 2014, p. 40) recommends averaging the adjusted person weights within a household and assigning the average weight to all household members. Such an averaging process ignores the individual response patterns in a household. In contrast, our alternative weighting approaches allow a nonresponse adjustment at the person level, without the need for subsequent averaging of the resulting weights. By including the common variables consistency is still ensured.

5 Efficiency Comparison of Person-Level and Integrated GREG Estimators

In Chapter 3, we derived in detail the consequences of the strict requirement of equal weights in the integrated weighting approach. Our MC simulation study strongly supports that these consequences result in more varied weights and coefficients as well as in less efficient point and variance estimates for small sample sizes. However, our deduced consequences and simulation results contradict Steel and Clark (2007) who claimed, in contrast, that the variance obtained by integrated weighting is less than the variance obtained by a person-level GREG estimator. Because of these contradictions, this chapter examines the theorems given by Steel and Clark (2007). Subsequently, we derive an efficiency comparison of both variances.

The remainder of this chapter is as follows: In Section 5.1, we reproduce the efficiency comparison given by Steel and Clark (2007). The main issues are, first, that they neglected the intercept in the integrated household-level model and thus, second, that the underlying assisting models are of different dimensions. Moreover, the interpretation of the difference between the variances using the argument of *controlling for* is ambiguous. Considering the discussed issue, in Section 5.2, we derive an own efficiency comparison. To be able to compare the variances of models of different dimensions, we initially desire to separate the effect of the intercept from the variance of an integrated GREG estimator. To solve this problem, we decompose the variance of an integrated GREG estimator into the variance of a reduced GREG estimator, which underlying model is of the same dimensions as the person-level GREG estimator, and add a constructed term that captures the effect of the intercept disregarded by the reduced model. Subsequently, we deduce a relationship between the coefficients of the person-level model and the household-level model to provide an interpretation of the efficiency comparison. At the end of this section, we are able to correctly compare the efficiency of the variances of a person-level and an integrated GREG estimator. The final outcome is explained by simulations results.

Finally, Section 5.3 suggests a further application field of the previously derived decomposition of regression coefficients to predict the difference between two coefficients of determination when adding or omitting explanatory variables. This further application can be relevant for econometricians as well as for survey statisticians. Section 5.4 summarizes the results of this chapter and draws conclusions.

To facilitate the reading of this chapter, we introduce some general indications. This chapter focuses on theoretical findings; therefore, all derivations given in the following refer to the population level instead to the sample. As such, whenever we use the term *variance*, we consider the variance $V(\hat{T}_y)$ instead of the estimated variance $\widehat{V}(\hat{T}_y)$. To better explain, we visualize our problems or solutions with graphs, such as Venn diagrams. Since the illustration via graphs is often limited to the case of having few variables, the proceeding in the following is twofold: Initially, we focus on the one-dimensional case and visualize the problems or solutions using graphs. Subsequently, we extend our findings to the multidimensional case. Our derived findings are summarized by results. Intermediate outcomes, essential in considering the results, are outlined within lemmas.

5.1 Efficiency Comparison Given by Steel and Clark (2007)

In Section 5.1.1, we reproduce two of the theorems derived by Steel and Clark (2007) considering the optimal estimator under single-stage cluster sampling and the difference between the contradictions of a person-level GREG estimator and an integrated GREG estimator. Subsequently, in Section 5.1.2, we detect two issues with their theorems. To underpin our argumentation, we present some results based on the same simulation setup as introduced in Section 3.4.1. Original text from Steel and Clark (2007) is indicated by boxes. For ease of understanding, we change the original notation into the notation of the present thesis.

5.1.1 Original Theorems Given by Steel and Clark (2007)

Steel and Clark (2007) compared the efficiency of a person-level and an integrated GREG estimator with households as basis. According to our Definition 3, the latter estimator is equivalent to an integrated GREG estimator with persons as a basis and $v_i = N_g^{-1}$ as variance parameter. However, in this chapter, we retain the term *integrated household-level* GREG estimator, as originally used by Steel and Clark (2007), for two reasons: a) the original denotation of the estimator at hand facilitates better comparability with the original paper and, b) the term *household-level* permits a more comprehensible interpretation of the neglected variable as an intercept, as we will see in the following.

In their first theorem, the optimal estimator for person characteristics under simple single-stage cluster sampling is derived. *Optimal* in this context means that the estimator has minimum variance in a large class of GREG estimators (for details see Section 2.3.3). Their first theorem is presented in the following box.

First Theorem Given by Steel and Clark (2007, p. 53): Optimal Estimator for Simple Cluster Sampling

Suppose that m households are selected by simple random sampling without replacement from a population of M households, and all people are selected from selected households. Consider the estimator of T_y given by

$$\hat{T}_y = \hat{T}_y^{\text{HT}} + \boldsymbol{h}^T(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})$$

where $\boldsymbol{h}$ is a constant vector of dimension Q. It is assumed that there exists a vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^T \boldsymbol{x_i} = 1$ for all $i \in U$. The variance of this estimator is minimized by $\boldsymbol{h}^*$ which are solutions of

$$\sum_{g \in s_h} (y_g - \boldsymbol{h}^T \boldsymbol{x_g})\boldsymbol{x_g} = \mathbf{0}$$

Hence $\hat{T}_y^{\text{INT}}$ with $v_i = N_g^{-1}$ for all $i \in U_p$ is the optimal choice of $\hat{T}_y$.

Note that $\hat{T}_y^{\text{INT}}$ is the integrated GREG estimator. From their first theorem, Steel and Clark (2007, p. 54) concluded that the variance of an integrated GREG estimator is less than or equal to that of a person-level GREG estimator and that the information discarded by summing up the original person-level information per household is irrelevant. Their theorem imply that the strict requirement of equal weights in the integrated weighting approach has no consequences. Moreover, the increased number of outcome values and the ignorance of the heterogeneity of the persons within a household, resulting in more spread weights and coefficients, would not affect the efficiency of the integrated GREG estimator. Even the one-to-one weight assignment between the levels, ignoring the different strengths of the relationship between the auxiliaries and the variable of interest, would cause any efficiency loss. These implications of their first theorem strongly contradict our simulation results of a comparison between an integrated and a person-level GREG estimator, comprehensively discussed in Section 3.4. Carrying their theorem too far, this would entail that under cluster sampling, it is always recommended to substitute the individual auxiliary information by the cluster-level aggregated information, independently from cluster size or within variance. The object of the following section is to detect some weaknesses on the theorem.

In their second theorem, Steel and Clark (2007) assessed the efficiency improvement of an integrated compared to a person-level GREG estimator by calculating the difference between both variances.

Second Theorem Given by Steel and Clark (2007, p. 54): Explaining the Difference in the Asymptotic Variances

Suppose that m households are selected by simple random sampling without replacement and all people are selected from selected households. Let $r_i^{B_p} = y_i - \boldsymbol{B_p}^T \boldsymbol{x_i}$ and let $\boldsymbol{B_c}$ be the result of regressing $r_i^{B_p}$ on $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ over $i \in U_p$ using weighted least squares regression weighted by N_g. Then

$$V(\hat{T}_y^{\text{GREG}}) - V(\hat{T}_y^{\text{INT}}) = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1}\boldsymbol{B_c}^T\Big(\sum_{g \in U_h} \boldsymbol{x_g}\boldsymbol{x_g}^T\Big)\boldsymbol{B_c} \tag{5.1}$$

where $\hat{T}_y^{\text{INT}}$ is calculated using $v_i = N_g^{-1}$ for all $i \in U_p$.

From the second theorem, Steel and Clark (2007, p. 54) concluded that the reduction in the variance from using an integrated household-level GREG estimator rather than a person-level GREG estimator is a quadratic form in $\boldsymbol{B_c}$. The discussion of this conclusion is given in detail in Section 5.1.2.2.

In the next section, we deeply discuss the presented theorems and the corresponding proofs. A detailed line-by-line discussion of the proofs of both theorems and further minor technical issues can be found in Section C.1 in Appendix C.

5.1.2 Issues of the Theorems Given by Steel and Clark (2007)

The main issues of the theorems are as followings:

- Steel and Clark (2007) tacitly assumed that the auxiliaries of a person-level GREG estimator sum up per household to the auxiliaries of an integrated GREG estimator and thus that both auxiliary vectors are of the same dimension. However, we show in Section 5.1.2.1 that the per-household summation of the person-level information results in a household-level auxiliary vector without an intercept.

- For the interpretation of their second theorem, Steel and Clark (2007) used the argument of *controlling for*, which clearly implies a multiple regression interpretation. However, we declare in Section 5.1.2.2 that their approach considerably differs from a multiple regression interpretation.

5.1.2.1 Neglecting the Intercept in the Integrated Household GREG Estimator

Steel and Clark (2007) assumed throughout their paper that the auxiliaries of a person-level GREG estimator sum up per household to the auxiliaries of an integrated household-level

GREG estimator. However, the summation per household of the person-level auxiliary vector of dimension Q

$$\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iQ})^T = (1, x_{i2}, \ldots, x_{iQ})^T$$

results in

$$\sum_{i \in U_g} \boldsymbol{x_i} = \boldsymbol{x_g} = (x_{g1}, x_{g2}, \ldots, x_{gQ})^T = (N_g, x_{g2}, \ldots, x_{gQ})^T. \tag{5.2}$$

Accordingly, $\boldsymbol{x_g}$ does not contain an intercept. In contrast, the person-level auxiliary vector of an integrated GREG estimator

$$\bar{\boldsymbol{x}}_{\boldsymbol{i}}^{\circ} = (\bar{x}_{i0}, \bar{x}_{i1}, \ldots, \bar{x}_{iQ})^T = (N_g^{-1},\ 1, \bar{x}_{i2}, \ldots, \bar{x}_{iQ})^T$$

is of dimension $(Q+1)$ and sums up per household to

$$\sum_{i \in U_g} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^{\circ} = \boldsymbol{x}_{\boldsymbol{g}}^{\circ} = (x_{g0}, x_{g1}, \ldots, x_{gQ})^T = (\ 1,\ N_g, x_{g2}, \ldots, x_{gQ})^T. \tag{5.3}$$

Therefore, $\boldsymbol{x}_{\boldsymbol{g}}^{\circ}$ contains an intercept, $x_{g0} = 1$, **and** the number of persons within a household, $x_{g1} = N_g$. Consequently, the auxiliaries of a person-level GREG estimator at the person level do not sum up to the auxiliaries of an integrated GREG estimator at the household level

$$\boldsymbol{x_g} \neq \boldsymbol{x}_{\boldsymbol{g}}^{\circ}.$$

It is important to note that we differentiate between a person-level intercept $x_{i1} = 1$, which sums up per household to $\sum_{i \in U_g} x_{i1} = N_g$, and a household-level intercept $x_{g0} = 1$.

The equality of $\boldsymbol{x_g}$ and $\boldsymbol{x}_{\boldsymbol{g}}^{\circ}$ is valid if and only if

1) the integrated auxiliary variables $\boldsymbol{x}_{\boldsymbol{g}}^{\circ}$ at the household level do not contain an intercept, or

2) the person-level auxiliary variables $\boldsymbol{x_i}$ contain N_g^{-1} as an additional auxiliary, which sums up to the intercept at the household level.

To point 1, including an intercept is crucial for several reasons. The first two reasons refer to survey statistics, the latter is originated in econometrics. Firstly, an intercept guarantees that the household weights sum up to the number of households in the population, M. Secondly, the sufficient condition $\sigma^2 = 1 = \boldsymbol{\lambda}^T \boldsymbol{x_g}$, guaranteeing the unbiasedness of a GREG estimator, is fulfilled for models comprising an intercept (cf. Särndal et al., 1989, p. 231). Thirdly, the residuals from a model without an intercept no longer sum up to zero. This affects the design-based variance formula, which has to be extended by the mean of the residuals.

To point 2, the additional auxiliary N_g^{-1} is required only in an integrated estimator to ensure the integrated property (see Section 3.1.1 for details). The integrated property induces that after a one-to-one weight assignment between both levels, the integrated person weights sum up to the number of persons within a household, and simultaneously the integrated weights at the household level sum up to the number of households. However, since the person-level weights are not assigned from the person to the household level, this variable is not required.

To conclude, no argumentation justifies point 1 or point 2 is fulfilled and thus that the equality of $\boldsymbol{x_g}$ and $\boldsymbol{x}_{\boldsymbol{g}}^{\circ}$ is valid. Instead, the auxiliary vectors of a person-level and an integrated GREG estimator are of different dimensions. Consequently, Steel and Clark (2007) neglected the intercept in the integrated GREG estimator.

At this point, the reason to use the name *integrated household-level* GREG estimator becomes more obvious. At the household level, the variable that determines the difference in dimension between the auxiliaries $\boldsymbol{x_g}$ and $\boldsymbol{x}_{\boldsymbol{g}}^{\circ}$ can be interpreted as household-level intercept $x_{g0} = 1$. In contrast, at the person level the counterpart of the household-level intercept is $x_{i0} = N_g^{-1}$, which has no clear interpretation. In the following, we derive the consequences of this misleading assumption for the previously presented theorems.

Consequence for Their First Theorem

In their first theorem, Steel and Clark (2007) derived the optimal estimator by differentiating the variance of a person-level GREG estimator with respect to the coefficient within the estimator. Certainly, the variance of a GREG estimator under simple single-stage cluster sampling

$$V(\hat{T}_y^{\text{GREG}}) = \frac{M^2}{m}\Big(1 - \frac{m}{M}\Big)(M-1)^{-1} \sum_{g \in U_h} \Big(\sum_{i \in U_g} y_i - \sum_{i \in U_g} \boldsymbol{x_i}^T \boldsymbol{b} \Big)^2$$

is minimized by

$$\boldsymbol{b}^* = \Big(\sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{x_g}^T \Big)^{-1} \sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{y_g}.$$

However, Steel and Clark (2007) drew a misleading conclusion for the optimality of the integrated household-level estimator because the following applies

$$\begin{aligned} \boldsymbol{b}^* &= (\sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{x_g}^T)^{-1} \sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{y_g} \\ &\neq (\sum_{g \in U_h} \boldsymbol{x}_{\boldsymbol{g}}^{\circ} \boldsymbol{x}_{\boldsymbol{g}}^{\circ T})^{-1} \sum_{g \in U_h} \boldsymbol{x}_{\boldsymbol{g}}^{\circ} \boldsymbol{y_g} = \boldsymbol{B}^{\circ}. \end{aligned}$$

Accordingly, the variance of a person-level GREG estimator is not minimized by the integrated coefficient $\boldsymbol{B}^{\circ}$, as claimed by Steel and Clark (2007). Nevertheless, it is surprising that the variance of a GREG estimator, constructed to estimate person characteristics, is optimized by a coefficient depending on the per-household aggregated auxiliary information $\boldsymbol{x_g}$ rather than on the individual person-level information $\boldsymbol{x_i}$. The surprise is reinforced, as we learned in Section 3.2.2 that the coefficients at the person and household level differ because of ecological fallacy. We come back to this paradox in Chapter 6 when discussing that the optimality is caused by the order of the sum and square root in the variance formula.

Consequence for Their Second Theorem
In their second theorem, Steel and Clark (2007) solved both variance formulas and combined them to one single term (see Equation 5.1). However, their mathematical rearrangements, when combining both formulas to derive (5.1), relies strongly on the assumption of the equality of $\boldsymbol{x}_g$ and $\boldsymbol{x}_g^\circ$. Otherwise, it would not be feasible to combine the auxiliaries of the variance of an integrated and a person-level GREG estimator into one single term. Consequently, Steel and Clark (2007) compared the efficiency of a person-level GREG estimator with that of a household-level GREG estimator without an intercept rather than with an integrated GREG estimator. We denote a household-level GREG estimator without an intercept as **reduced household-level model**, hereinafter.

To underpin that there is a distinction between, on the one hand comparing a person-level GREG estimator with an integrated GREG estimator, as originally intended by Steel and Clark (2007),

$$V(\hat{T}_y^{\text{GREG}}) - V(\hat{T}_y^{\text{INT}}), \tag{5.4}$$

and on the other hand comparing a person-level GREG estimator with a reduced household-level GREG estimator, as actually realized by Steel and Clark (2007),

$$V(\hat{T}_y^{\text{GREG}}) - V(\hat{T}_y^{\text{SC}}), \tag{5.5}$$

we compute the densities for the correct difference (5.4) and for the difference indicated by Steel and Clark (2007) (5.5) for $R = 1000$ MC replications. The densities are computed based on the AMELIA data set within the same simulation setting introduced in Section 3.4.1. Superscript SC refers to Steel and Clark. It should be remarked that the plots are given on different scales, since we are interested in the comparison of both approaches and not a comparison of the different variables of interest.

We learn from Figure 5.1 that in particular for variables related to the household, Steel and Clark (2007) underestimate the correct difference between the variances of an integrated and a person-level GREG estimator. Figure 5.2 plots the average household size within each MC replication against the deviation between the correct approach and the approach by Steel and Clark (2007). For this plot, we choose four variables of interest presented in Figure 5.1. `inc_hs5` and `inc_hs6` are characterized by a large deviation between the two approaches. `bene_age4` and `inc` are characterized by a small deviation between the two approaches. It becomes evident that for `inc_hs5` and `inc_hs6` the deviation increases with the average household size. In contrast, for `bene_age4` and `inc`, both measures seems to be unrelated. Hence, the amount of the deviation between including (correct approach) and not including (Steel and Clark approach) an intercept in the household model depends on the average household size in the sample. In other words, if large households prevail in the sample, Steel and Clark (2007) underestimate the correct variance of an integrated GREG estimator.

5.1.2.2 Interpretation of the Difference between the Asymptotic Variances

To focus on the interpretation of the second theorem, we temporarily assume that either point 1) or point 2) in Section 5.1.2.1 is fulfilled and thus that the assumption of the equality of $\boldsymbol{x}_g$ and

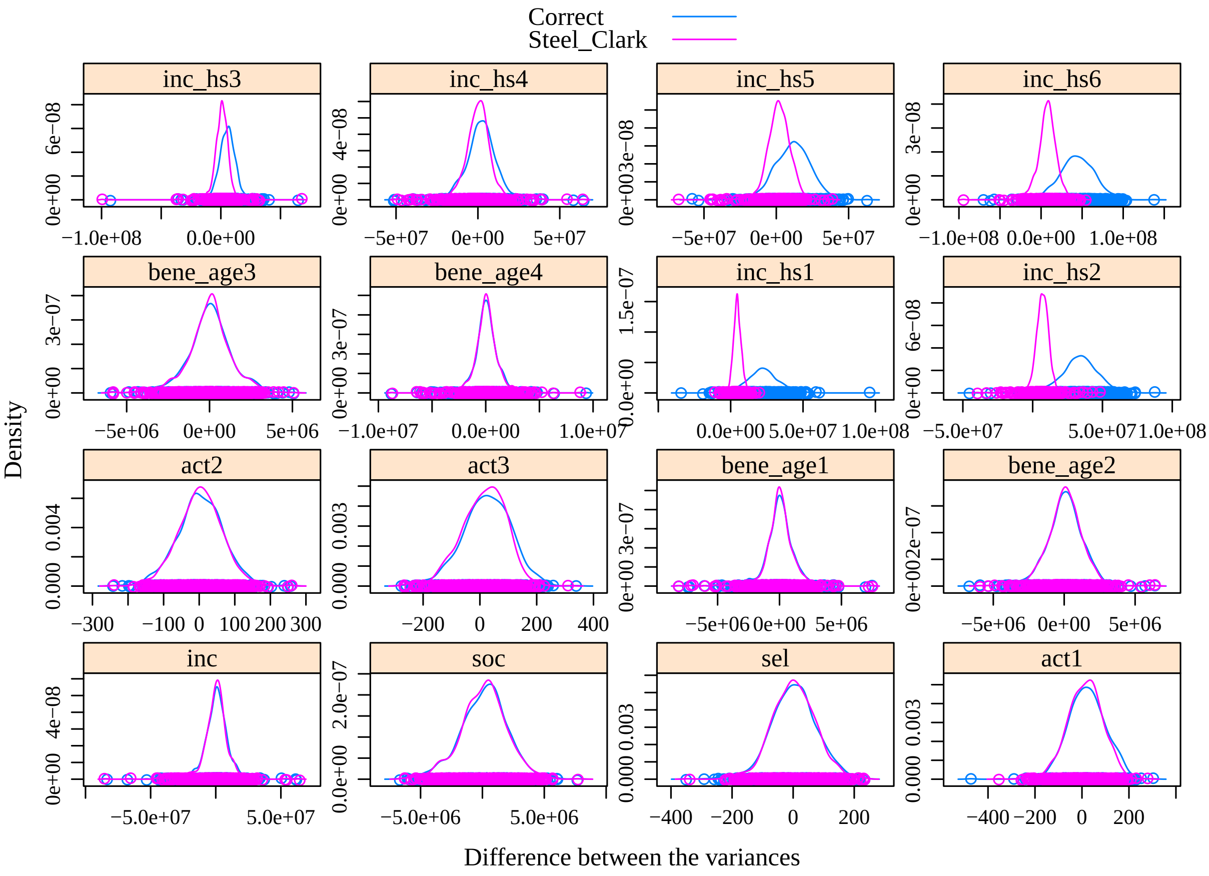

Figure 5.1: Density plots for the difference between the variances according to the correct approach and the approach of Steel and Clark (2007) for $m = 1500$

$\boldsymbol{x}_{\boldsymbol{g}}^{\circ}$ is valid. We skip this assumption in the next Section 5.2 when providing a correct efficiency comparison.

The difference between the variances in (5.1) derived by Steel and Clark (2007) depends on the coefficient

$$\boldsymbol{B_c} = \left(\sum_{i \in U_p} N_g \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}{}^T \right)^{-1} \sum_{i \in U_p} N_g \bar{\boldsymbol{x}}_{\boldsymbol{i}} r_i^{B_p}, \tag{5.6}$$

which results from regressing $r_i^{B_p} = y_i - \boldsymbol{B_p}^T \boldsymbol{x_i}$ on $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ using GLS. From their second theorem, Steel and Clark (2007, p. 54) argued that: “The result shows that the reduction in variance from using $\hat{T}_y^{\text{INT}}$ $(v_i = N_g^{-1})$ rather than $\hat{T}_y^{\text{GREG}}$ is a quadratic form in $\boldsymbol{B_c}$. Hence the extent of the improvement depends on the extent to which $\bar{\boldsymbol{x}}_i$ helps to predict y_i after $\boldsymbol{x_i}$ has already been controlled for, i.e., the extent to which a linear contextual effect helps to predict $r_i^{B_c}$ over $i \in U_p$, using a weighted least squares regression weighted by N_g.“[1]

[1] For a better understanding of the statement, we change the original notation into the notation of the present thesis.

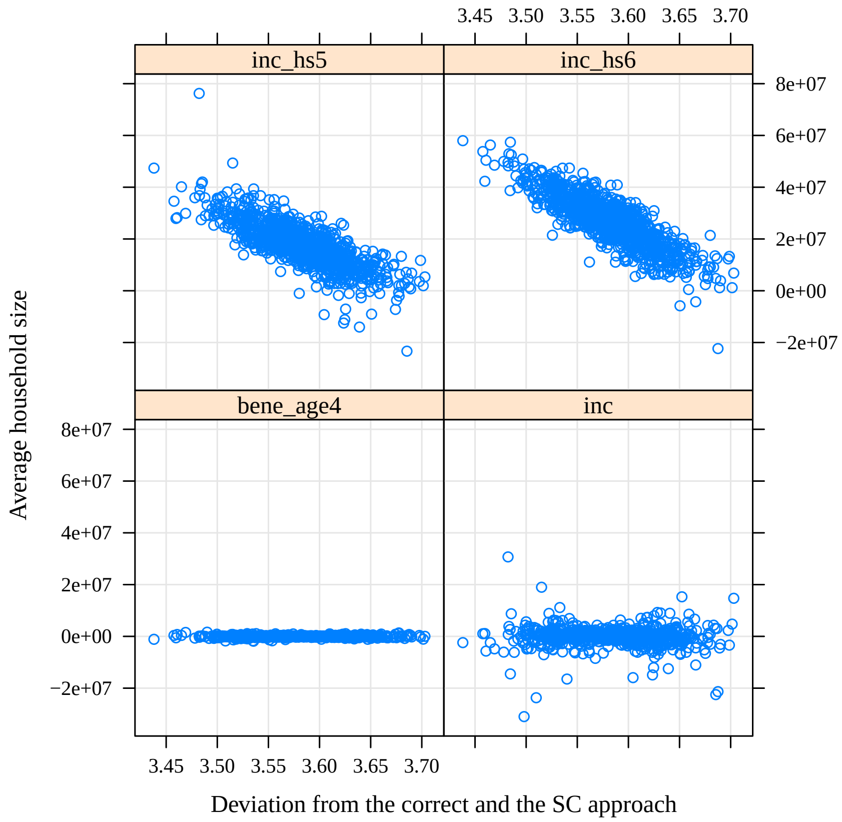

Figure 5.2: Scatterplots for the deviation between the correct approach and the approach of Steel and Clark (2007) for $m = 1500$

Using the argument *controlling for* is a clear hint for a multiple regression interpretation. In a multiple regression, the interpretation is conducted *ceteris paribus*, meaning the coefficient of a certain explanatory variable describes its effect on the variable of interest holding all other explanatory variables constant (cf. Wooldridge, 2013, p. 70). However, we will show that $\boldsymbol{B}_c$ does not describe the extent to which $\bar{\boldsymbol{x}}_i$ helps to predict y_i after $\boldsymbol{x}_i$ has already been controlled for. To emphasize this presumption, we apply the Frisch-Waugh-Lovell theorem (cf. Frisch and Waugh, 1933; Lovell, 1963) and Venn diagrams. Typically, Venn diagrams address the case of having one or two explanatory variables. In the case of three explanatory variables a simplex representation would be needed to draw Venn diagrams. Hence, we initially focus on the simple case of two explanatory variables in order to visualize the ambiguous interpretation of $\boldsymbol{B}_c$ with Venn diagrams. Subsequently, we extend our findings to the multiple variable case.

The **Frisch-Waugh-Lovell (FWL) theorem** is helpful for declaring the meaning of *controlling for*. The FWL theorem states that in a multiple regression the coefficient of any single variable can also be obtained by first partialing out the effects of all other explanatory variables from both the specific single variable and the variable of interest, and then regressing the remaining variation of the variable of interest on the remaining variation of the explanatory variables. For explanation, consider the following model,

$$y_i = H_x x_i + H_{\bar{x}} \bar{x}_i + r_i^H, \tag{5.7}$$

introduced by Steel and Clark (2007) to motivate the interpretation of their second theorem. Applying the FWL theorem to $H_{\bar{x}}$, we obtain the following two regressions,

$$\begin{aligned} y_i &= B_p x_i + r_i^{B_p}, \quad \text{and} \\ \bar{x}_i &= B_x x_i + r_i^{B_x}, \end{aligned} \tag{5.8}$$

by partialling out the effect of x_i from both y_i and $\bar{x}_i$. Then, by regressing the remaining variation of the variable of interest, captured by $r_i^{B_p}$, on the remaning variarion of the explanatory variable, captured by $r_i^{B_x}$, given by

$$r_i^{B_p} = H_{\bar{x}} r_i^{B_x} + r_i^{H_{\bar{x}}}, \tag{5.9}$$

the coefficient $H_{\bar{x}}$ from the initial regression in (5.7) results. According to the *ceteris paribus* interpretation, the coefficient $H_{\bar{x}}$ describes the effect of $\bar{x}_i$ on y_i controlled for any effect of x_i. It should be remarked that the residual r_i^H in (5.9) exactly conforms with the residual from model (5.7).

Venn diagrams help to visualize the difference between the meaning of *controlled for* and the interpretation claimed by Steel and Clark (2007). We refer to the interpretation of Venn diagrams as suggested by Kennedy (1981, 2002) in the context of regression analysis.[2] Figure 5.3 illustrates the multiple regression model (5.7). Each circle represents the variation of a variable. Intersections between two variables are interpreted as variation common to both variables. The common variation of a variable of interest and an explanatory variable determines the information used for estimating the corresponding regression coefficient. The green shaded area between the circles y_i and $\bar{x}_i$, for example, describes the information used to calculate the coefficient $H_{\bar{x}}$. The variation common to the variables x_i and $\bar{x}_i$ (not shaded) cannot be clearly assigned to one variable and is thus are not used for calculating $H_{\bar{x}}$. The remaining variation of the variables, outside the intersections, determines the residuals of a regression.

Figure 5.4 visualizes the FWL theorem. According to this, $H_{\bar{x}}$ obtained from the multiple regression (5.7), can alternatively be calculated by regressing the residuals r_i^B (yellow shaded) on the residuals $r_i^{B_x}$ (blue shaded).

In contrast to this, Venn diagram 5.5 illustrates the approach given by Steel and Clark (2007). They skipped the second regression in (5.8) and regressed $r_i^{B_p}$ on $\bar{x}_i$ which is given by

$$r_i^{B_p} = B_c \bar{x}_i + r_i^{B_c} \tag{5.10}$$

instead of $r_i^{B_p}$ on $r_i^{B_x}$ given by (5.9) as it would be correct following the FWL theorem. Therefore, Steel and Clark (2007) did to partial out the effect of $\bar{x}_i$ on x_i (second regression in (5.8)). Nevertheless, for interpretation of B_c, they used the argument *controlling for*. A comparison

[2] Different concepts for interpreting Venn diagrams in regression analysis exist. In contrast to Cohen et al. (2013) and Ip (2001), it is not Kennedy's attempt to exposit R^2.

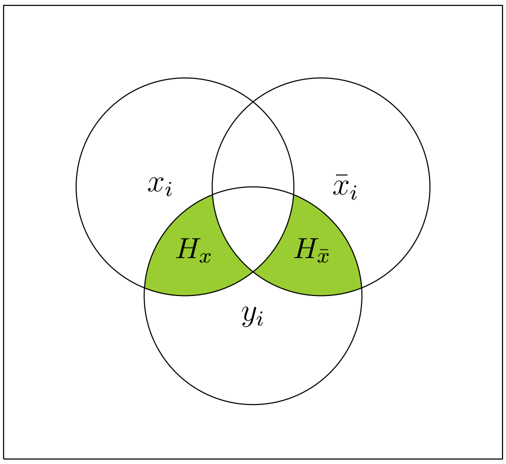

Figure 5.3: Venn diagram for a simple contextual model

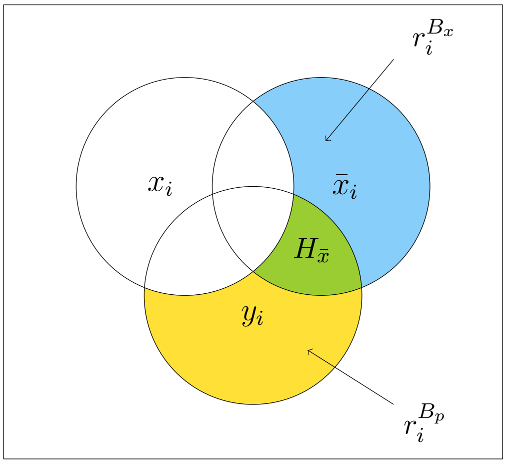

Figure 5.4: Venn diagram illustrating the Frisch-Waugh-Lovell theorem

of Figures 5.4 and 5.5 makes evident that the difference between term *controlling for* and the approach by Steel and Clark (2007), and thus the difference between $H_{\bar{x}}$ and B_c, is quantified by the intersection of x_i and $\bar{x}_i$ (red shaded in Figure 5.5). Thus, for calculating B_c the complete variation of $\bar{x}_i$ is used, including the variation common to $\bar{x}_i$ and x_i. As result, $H_{\bar{x}}$, as the coefficient describing the effect of x_i on y_i *controlled for* x_i, and B_c, for which Steel and Clark (2007) use this interpretation. The difference can be formalized by

$$\begin{aligned} H_{\bar{x}} &= \frac{\sum_{i \in U_p} N_g r_i^{B_x} r_i^{B_p}}{\sum_{i \in U_p} r_i^{B_x} r_i^{B_x}} \\ &\neq \frac{\sum_{i \in U_p} N_g \bar{x}_i r_i^{B_p}}{\sum_{i \in U_p} \bar{x}_i^2} \\ &= B_c. \end{aligned} \tag{5.11}$$

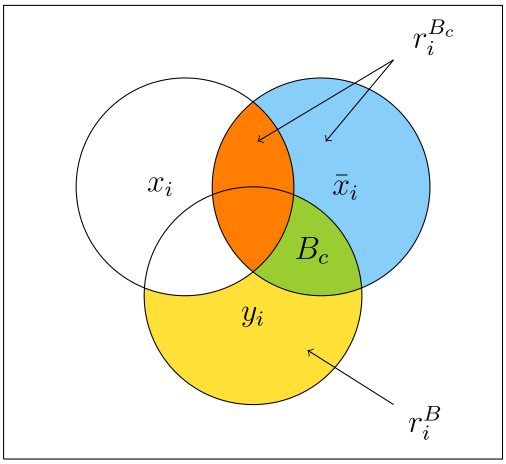

Figure 5.5: Venn diagram illustrating the approach of Steel and Clark (2007)

The inequality is valid, as long as the vectors $\boldsymbol{x} = (x_1, \dots, x_N)^T$ and $\bar{\boldsymbol{x}} = (\bar{x}_1, \dots, \bar{x}_N)^T$ are not orthogonal. If $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ are orthogonal, than there would be no common variation between the two auxiliaries and the intersection between both circles in the Venn diagram would vanish. Therefore, $r_i^{B_x} = \bar{x}_i$ and $H_{\bar{x}} = B_c$. However, the case of $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ are orthogonal is very unlikely, because $\bar{x}_i$, as the household mean value, is a function of x_i.

Remark 4. *Even if $B_c \neq H_{\bar{x}}$ is valid, their corresponding shaded areas in Figures 5.4 and 5.5 seem to be of the same magnitude. Certainly, in a Venn diagram the intersection between two variables is interpreted as common information used to calculate the regression coefficient. Although both areas of B_c and $H_{\bar{x}}$ are of the same magnitude, their corresponding numeric values differ as their independent variables differ: $\boldsymbol{r}^{\boldsymbol{B_x}} \neq \bar{\boldsymbol{x}}$. The only reason for the same amount of common variation between y_i and $r_i^{\bar{x}}$ as well as between y_i and $\bar{x}_i$ is that in both regressions the dependent variable is the same. In contrast, the magnitude of the area beyond the intersection reflects the magnitude of a parameter estimate, which cannot be explained by the regressors (cf. Kennedy, 2002).*

The described inequality of the coefficients is also valid for vectors containing $Q > 2$ auxiliary variables with $\boldsymbol{x_i} = (x_{i1}, \dots, x_{iQ})^T$ and $\bar{\boldsymbol{x}}_{\boldsymbol{i}} = (\bar{x}_{i1}, \dots, \bar{x}_{iQ})^T$. In the multiple case, inequality (5.11) is demonstrated by

$$\begin{aligned}
\boldsymbol{H}_{\bar{\boldsymbol{x}}} &= \left(\sum_{i \in U_p} N_g \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{B_x}} \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{B_x} T} \right)^{-1} \sum_{i \in U_p} N_g \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{B_x}} r_i^{B_p} \\
&\neq \left(\sum_{i \in U_p} N_g \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^T \right)^{-1} \sum_{i \in U_p} N_g \bar{\boldsymbol{x}}_{\boldsymbol{g}} r_i^{B_p} \\
&= \boldsymbol{B_c}.
\end{aligned}$$

To conclude, although Steel and Clark (2007) used a different approach, they justified their interpretation using the argument of *controlling for*. The difference between *controlling for* and the approach of Steel and Clark (2007) is quantified by the common variation of $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_i$ determined by the intersection in Figure 5.5 (shaded in red). The magnitude of the intersection depends on the household sizes: the smaller the households, the higher the correlation between the two variables, the larger the intersection, and the greater the difference between *controlling for* and the approach of Steel and Clark (2007).

5.2 Efficiency Comparison of a Person-Level GREG estimator and an Integrated Household-Level GREG Estimator

After elaborating the issues of the efficiency comparison given by Steel and Clark (2007), the aim of this section is to provide a correct comparison of the variances of a person-level GREG estimator and an integrated GREG estimator considering that the auxiliaries are of different dimensions.

First, to be able to compare the variances of models of different dimensions, we separate in Section 5.2.1 the effect of the intercept from the variance of an integrated household-level GREG estimator. To solve this problem, we decompose the variance of an integrated household-level GREG estimator into the variance of a reduced household-level GREG estimator, which underlying model is of the same dimension as the person-level GREG estimator, and construct a term that captures the effect of the intercept disregarded by the reduced household-level model. The decomposition allows us to compare the variances of models of different dimensions. By inserting the decomposition into the efficiency comparison we obtain an intermediate result (Section 5.2.2).

Second, as we doubt the appropriateness of the interpretation of *controlling for*, the objective of Section 5.2.3 is to derive a functional relationship between the coefficient from the reduced household-level and the person-level GREG estimator, which are of the same dimension. To solve this problem, we apply a model that simultaneously contains both the auxiliaries of the person-level and the reduced household-level model. By inserting the functional relationship into the intermediate result, derived before, the final result of our efficiency comparison results (Section 5.2.4).

It is important to note that even if we are interested in an efficiency comparison of a person-level and an integrated GREG estimator for person characteristics, we have to consider both auxiliary variable sets at the household level, because the variance formula under cluster sampling refers to the aggregates of the variables (see Section 2.4). Therefore, instead of determining the auxiliary vectors $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_i^{\circ}$, we compare their aggregated values $\boldsymbol{x}_g$ and $\boldsymbol{x}_g^{\circ}$.

For a better understanding, we deduce the proofs in detail in the following sections. To verify the correctness of the mathematical rearrangements in the proofs, we program every line as

R code. For that, we draw a sample of $m = 1500$ households by means of simple random sampling from the AMELIA data set.

5.2.1 Separating the Effect of the Intercept from the Variance of an Integrated Household-Level GREG Estimator

The issue of correctly computing the difference of the variances between a person-level and an integrated GREG estimator can be formalized by the following objective function

$$\begin{aligned}\left(\mathrm{V}(\hat{T}_y^{\mathrm{GREG}}) - \mathrm{V}(\hat{T}_y^{\mathrm{INT}})\right) \Big/ \frac{M^2}{m}\left(1-\frac{m}{M}\right)(M-1)^{-1} \\ = \sum_{g\in U_h}\left(r_g^{B_p}\right)^2 - \sum_{g\in U_h}\left(r_g^{B_h^\circ}\right)^2, \end{aligned} \tag{5.12}$$

where the residuals are obtained from a person-level model and an integrated household model

$$y_i = \boldsymbol{B_p}^T \boldsymbol{x_i} + r_i^{B_p} \quad \text{and} \quad y_g = \boldsymbol{B_h^{\circ}}^T \boldsymbol{x_g^\circ} + r_g^{B_h^\circ},$$

respectively. We divide the objective function (5.12) by the term $\frac{M^2}{m}(1-\frac{m}{M})(M-1)^{-1}$, as under simple single-stage cluster sampling (see Section 2.4) it emerges in both variance formulas. Thereby, we can neglect this term in the following.

Unfortunately, the auxiliaries $\boldsymbol{x_i}$ and $\boldsymbol{x_g^\circ}$ of a person-level and integrated GREG estimator are of different dimensions. The difference is constituted by the household-level intercept x_{g0} (see Equations (5.2) and (5.3)). Therefore, to provide a correct efficiency comparison, we aim at separating the intercept from the integrated household-level model. This issue is summarized by the following problem.

Problem 1. ***Separating the Effect of a Variable from the Variance***
We aim at separating the effect of a variable from the variance and constructing a term that captures the disregarded effect of the variable on the initial variance.

To solve Problem 1, we decompose the sum of squared residuals of an integrated household-level GREG estimator into the sum of squared residuals of a reduced-household-level GREG estimator, whose underlying model is of the same dimension as the person-level GREG estimator. We then construct a remaining term to capture the effect of the intercept we disregarded previously. Translated into our objective (5.12), that means

$$\begin{aligned}\left(\mathrm{V}(\hat{T}_y^{\mathrm{GREG}} - \mathrm{V}(\hat{T}_y^{\mathrm{INT}})\right) \Big/ \frac{M^2}{m}(1-\frac{m}{M})(M-1)^{-1} \\ = \sum_{g\in U_h}(r_g^{B_p})^2 - \sum_{g\in U_h}(r_g^{B_h^\circ})^2 \\ = \sum_{g\in U_h}(r_g^{B_p})^2 - \sum_{g\in U_h}(r_g^{B_h})^2 + \text{remaining term}, \end{aligned} \tag{5.13}$$

where $r_g^{B_h}$ is the residual of model

$$y_g = \boldsymbol{B_h}^T \boldsymbol{x_g} + r_g^{B_h}, \tag{5.14}$$

which we denote hereinafter as **reduced household-level model**. It does not contain an intercept. According to (5.13), we have to quantify the remaining term.

The solution to Problem 1 consists of three steps:

- We start in Section 5.2.1.1 by decomposing the coefficient from an integrated household-level model into a coefficient resulting from a reduced household model omitting an intercept and a remaining factor that captures the effect of the intercept on the integrated household-level coefficient that is disregarded by the reduced household-level model.

- We continue in Section 5.2.1.2 with translating the decomposition of the integrated coefficient to the corresponding residuals, since we are initially interested in the separation of the intercept from the variance.

- Finally, in Section 5.2.1.3, we extend our findings from the previous sections to the sum of squared residuals, which is not straightforward as squaring is a non-linear transformation.

At the end of this section, we are able to correctly compare the variances of any GREG estimators containing different numbers of auxiliaries.

5.2.1.1 Decomposition of the Coefficients of an Integrated Household-level Model

We start by decomposing the integrated coefficient $\boldsymbol{B_h^\circ}$ into a coefficient $\boldsymbol{B_h}$ resulting from a reduced household model without an intercept (5.14) and additionally a remaining term. The decomposition of the integrated coefficient is impressive, as in a multiple regression the coefficients are calculated *ceteris paribus* and therefore incorporate the covariances between the variables.

For a better comprehension, we visualize our separation problem with Venn diagrams. Therefore, as done before, we will initially focus i) on simple models comprising an intercept and one auxiliary variable. Subsequently, we extend our findings to ii) multiple models comprising an intercept and $Q > 1$ auxiliary variables.

i) Simple Models Comprising an Intercept and a Single Auxiliary Variable
Consider x_{i1} as an auxiliary variable at the person level which sum up to $x_{g1} = \sum_{i \in U_g} x_{i1}$, the auxiliary variable of the reduced household-level model. The integrated household-level auxiliaries additionally contain an intercept: $\boldsymbol{x_g^\circ} = (x_{g0}, x_{g1})^T = (1, x_{g1})^T$.

The integrated household-level model in the simple case can be expressed as

$$\begin{aligned} y_g &= \boldsymbol{B_h^\circ}^T \boldsymbol{x_g^\circ} + r_g^{B_h^\circ} \\ &= B_{x_0}^\circ x_{g0} + B_{x_1}^\circ x_{g1} + r_g^{B^\circ} \end{aligned} \tag{5.15}$$

with $\boldsymbol{B}_{\boldsymbol{h}}^{\circ} = (B_{x_0}^{\circ}, B_{x_1}^{\circ})^T$ as coefficient vector. Omitting the intercept x_{g0} results in the one-dimensional reduced household-level model

$$y_g = B_h x_{g1} + r_g^{B_h}. \tag{5.16}$$

It should be noted that models (5.15) and (5.16) both explain the same variable of interest y_g. The coefficients $B_{x_1}^{\circ}$ and B_h refer to auxiliary x_{g1}, which is common to both models. Their relation is illustrated by Venn diagrams 5.6 and 5.7[3], with $B_{x_1}^{\circ}$ and B_h highlighted in blue. $B_{x_1}^{\circ}$ is a coefficient from the multiple regression model (5.15). This means that it captures the effect of x_{g1} on y_g controlled for the effect of x_{g0}. In contrast, B_h describes the effect of the same x_g on y_g not controlled for x_{g0}. Hence, the green shaded area in Figure 5.7, describing the information used to calculate B_h, exceeds the yellow shaded area in Figure 5.6, which in turn is used to calculate $B_{x_1}^{\circ}$.

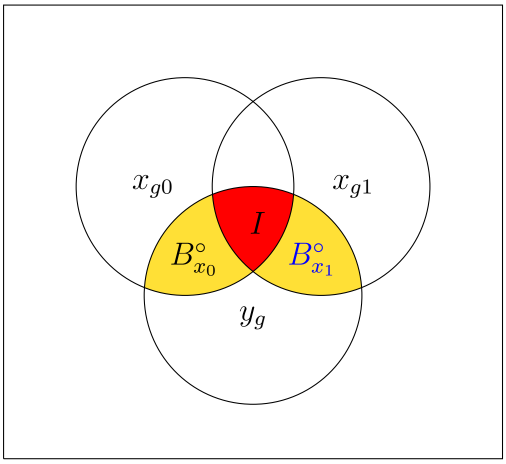

Figure 5.6: Venn diagram illustrating the integrated coefficient $\boldsymbol{B}_{\boldsymbol{h}}^{\circ} = (B_{x_0}^{\circ}, B_{x_1}^{\circ})^T$

Now, in order to separate the effect of the intercept x_{g0} from the integrated household model, we aim to decompose the integrated coefficients $\boldsymbol{B}_{\boldsymbol{h}}^{\circ} = (B_{x_0}^{\circ}, B_{x_1}^{\circ})^T$, shaded yellow in Figure 5.6, into the following:

- a coefficient $B_{x_0}^{\circ}$ (shaded gray in Figure 5.7) describing the effect of the intercept x_{g0} on y_g controlling for x_{g1},
- a coefficient B_h (shaded green in Figure 5.7) from the reduced household-level model (5.16) excluding an intercept. It therefore equals in dimension x_{i1}, the auxiliaries of the person-level model, and

[3] In contrast to the Venn diagrams shown previously, the intercept x_{g0}, as an independent variable, does not vary. Thus, the circle of x_{g0} cannot be interpreted as variation of x_{g0}. Nevertheless, one can calculate its corresponding coefficient as we will derive in the following. Therefore, we still use Venn diagrams to illustrate the decomposition, even though x_{g0} is constant.

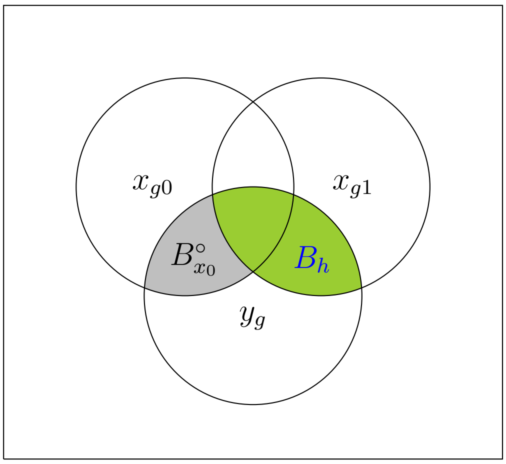

Figure 5.7: Venn diagram illustrating the reduced household-level coefficient B_h

- a remaining factor capturing the common variation of x_{g1} and the intercept x_{g0} when calculating $\boldsymbol{B}^{\circ}_{\boldsymbol{h}}$ quantified by the intersection I between the circles of y_g, x_{g0} and x_{g1} (shaded red in Figure 5.6).

Following Figures 5.6 and 5.7, we can rewrite B°_{x1} as the difference between B_h and the intersection I between the circles of y_g, x_{g1} and x_{g0}. Consequently, the decomposed household-model coefficient is given by

$$\boldsymbol{B}^{\circ}_{\boldsymbol{h}} = \begin{pmatrix} B^{\circ}_{x_0} \\ B^{\circ}_{x_1} \end{pmatrix} = \begin{pmatrix} B^{\circ}_{x_0} \\ B_h - \text{intersection } I \end{pmatrix} . \tag{5.17}$$

Accordingly, we need to quantify the intersection I (shaded in red in Venn diagram 5.7). Unfortunately, a simple decomposition of $\boldsymbol{B}^{\circ}_{\boldsymbol{h}}$ into $B^{\circ}_{x_0}$ and $B^{\circ}_{x_1}$ and the FWL theorem does not offer a solution. The reason is that $B^{\circ}_{x_0}$ and $B^{\circ}_{x_1}$ result from one common multiple regression and thus are calculated as partial coefficients. However, we are interested in the residuals from separated regressions, of which at least one regression excludes the intercept and thus has the same dimension as the person-level regression.

Fortunately, the **mediation model**, known from psychology and sociology, provides a promising solution for quantifying the intersection I. Within this framework, it is assumed that the relation between a variable of interest and an explanatory variable is more complex than a directly observed bivariate relation. Instead the explanatory variable may be intervened by a non-observable so-called mediator variable which in turn influences the variable of interest (cf. MacKinnon et al., 2007; MacKinnon, 2008). Mediator variables are also known in the literature as intermediary variables, intervening variables, suppressors, covariates, or moderators. Figure 5.8 illustrates how the direct, observable relation between a variable of interest y_g and an explanatory variable x_g is intervened by a mediator variable m_g. The mediator variable m_g simultaneously represents a variable of interest (in relation to x_g) and an explanatory variable

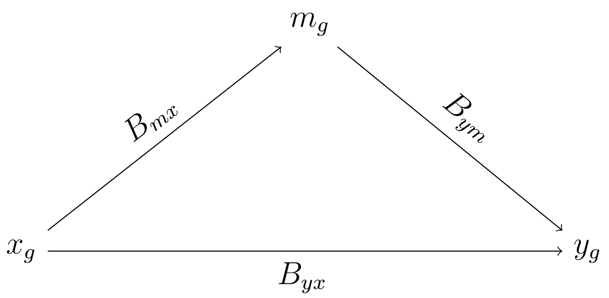

Figure 5.8: Total, direct and indirect effects in a mediation model

(in relation to y_g). For a detailed discussion, the interested reader is referred to Judd and Kenny (1981), Frazier et al. (2004), and Fairchild and MacKinnon (2009).

Baron and Kenny (1986) proposed to differentiate between total, direct, and indirect effects. **Direct effects**, here B_{mx}, B_{ym} and B_{yx}, cannot be intervened via third variables and arise from the following regressions

$$\begin{aligned} m_g &= B_{mx} x_g + r_g^m \\ y_g &= B_{yx} x_g + B_{ym} m_g + r_g^y . \end{aligned}$$

The direct effects B_{mx} and B_{ym} constitute the **indirect effect** of x_g on y_g via m_g

$$B_{mx \cdot ym}^{indirect} = B_{mx} \cdot B_{ym} .$$

Then, the **total effect** resulting from a bivariate regression of x_g on y_g can be split into a direct effect of x_g on y_g controlling for the effect of m_g and an indirect effect of x_g via m_g

$$B_{yx}^{total} = B_{yx} + B_{mx \cdot ym}^{indirect} . \tag{5.18}$$

Even though we are interested in neither indirect effects nor in mediator models per se, we can translate the splitting process into total, direct, and indirect effects to quantify the intersection I. Applying the splitting process to our problem of the decomposition of $\boldsymbol{B}_h^\circ$, the intercept x_{g0} can be interpreted as mediator variable (Figure 5.9). The coefficient B_h of the reduced household model (5.16) represents the total effect resulting from regressing y_g solely on x_{g1} (without an intercept). The direct effects B_{x0}° and B_{x1}° are partial regression coefficients of the integrated household-level model. The third direct effect F_{x_1} arises from a model with the intercept x_{g0} as variable of interest

$$x_{g0} = F_{x_1} x_{g1} + r_g^{F_{x_1}} . \tag{5.19}$$

We denote model (5.19) as the **auxiliary model**. The denotation *auxiliary* emphasizes that the only purpose of the auxiliary model is to deliver F_{x_1}, which is required to determine the indirect effect. We will refer to this kind of model below.

$$\begin{tikzcd}
 & x_{g0} \arrow[dr, "B^\circ_{x_0}"] & \\
x_{g1} \arrow[ur, "F_{x_1}"] \arrow[rr, "B^\circ_{x_1}"'] & & y_g
\end{tikzcd}$$

Figure 5.9: Total, direct and indirect effects applied to the decomposition of $\boldsymbol{B_h^\circ}$

In line with the mediation model and the equation in (5.18) we can rewrite B_h as

$$B_h = B^\circ_{x_1} + \underbrace{B^\circ_{x_0} F_{x_1}}_{\text{intersection } I} . \tag{5.20}$$

As a result, the second term on the right-hand side in (5.20) exactly defines the intersection I we need to decompose $\boldsymbol{B_h^\circ}$. Thus, its magnitude depends on the relationship between x_{g0} and x_{g1} and on the effect of x_{g0} on y_g controlled for x_{g1}.

When intersection I from (5.20) is inserted into (5.17), the decomposed household-level coefficient immediately results

$$\boldsymbol{B_h^\circ} = \begin{pmatrix} B^\circ_{x_0} \\ B^\circ_{x_1} \end{pmatrix} = \begin{pmatrix} B^\circ_{x_0} \\ B_h - B^\circ_{x_0} F_{x_1} \end{pmatrix} . \tag{5.21}$$

In the following, we briefly discuss how the coefficients of a reduced and an auxiliary model differ compared to coefficients obtained from ordinary regressions.

Coefficients from the Reduced Household Model. The reduced household-level model (5.16) omits an intercept. This implies that the regression line runs through the origin. Econometrically speaking, a regression line through the origin means that when all auxiliaries are set to zero, the expected value of the variable of interest also equals zero. Consequently, the interpretation of the slope parameter, as remaining coefficient, changes. To clarify the difference between the coefficients from models with and without an intercept, we derive the formulas. We start with B_h as coefficient from the reduced household model (5.16). Straightforward from the least squares theory (cf. Greene, 2003, Section 6.4; Wooldridge, 2013, Section 3.2) B_h is determined by minimizing the sum of squared residuals

$$\min_{B_h} \sum_{g \in U_h} (r_g^{B_h})^2 = \sum_{g \in U_h} (y_g - B_h x_{g1})^2 .$$

Setting the first derivative of the minimization problem with respect to B_h equal to zero, we obtain

$$\begin{aligned}\frac{\partial \sum_{g\in U_h}(r_g^{B_h})^2}{\partial B_h} &= 2\sum_{g\in U_h}(y_g - \ B_h x_{g1})(-x_{g1}) \overset{!}{=} 0\\ \Leftrightarrow \qquad \sum_{g\in U_h} y_g x_{g1} &= B_h \sum_{g\in U_h} x_{g1}^2\\ \Leftrightarrow \qquad B_h &= \frac{\sum_{g\in U_h} y_g x_{g1}}{\sum_{g\in U_h} x_{g1}^2}.\end{aligned}$$

In contrast, the coefficient from an integrated household model $y_g = B^\circ_{x_0} x_{g0} + B^\circ_{x_1} x_{g1} + r_g^{B^\circ_h}$, as an ordinary model comprising an intercept, equals

$$B^\circ_{x_1} = \frac{\sum_{g\in U_h}(y_g - \bar{y})(x_{g1} - \bar{x}_1)}{\sum_{g\in U_h}(x_{g1} - \bar{x}_1)^2}$$

with $\bar{y} = M^{-1}\sum_{g\in U_h} y_g$ and $\bar{x}_1 = M^{-1}\sum_{g\in U_h} x_{g1}$ as mean values. Comparing the formulas of B_h and $B^\circ_{x_1}$, both referring to the effect of x_1 on y_g, it becomes apparent that the former coefficient no longer determines the covariance, as it does not include mean values.

Coefficients from the Auxiliary Model. The coefficient from the auxiliary model (5.19) also differs from the coefficient from an ordinary regression, as it arises from regressing on a constant variable of interest and additionally it does not contain an intercept. Setting the first derivative of the following minimizing problem

$$\min_{F_x} \sum_{g\in U_h}(r_g^{F_{x_1}})^2 = \sum_{g\in U_h}(x_{g0} - F_{x_1} x_{g1})^2$$

equal to zero, we obtain

$$\begin{aligned}\frac{\partial \sum_{g\in U_h}(r_g^{F_{x_1}})^2}{\partial F_{x_1}} &= 2\sum_{g\in U_h}(x_{g0} - \ F_{x_1} x_{g1})(-x_{g1}) \overset{!}{=} 0\\ \Leftrightarrow \qquad F_{x_1} &= \frac{\sum_{g\in U_h} x_{g1}}{\sum_{g\in U_h} x_{g1}^2}.\end{aligned}$$

Therefore, F_{x_1} no longer depends on the dependent variable.

However, the only aim of the reduced household and the auxiliary model is to decompose the integrated household-level coefficient $\boldsymbol{B}^\circ_h$. Hence, we are not interested in the interpretation of the coefficients B_h and F_{x_1} per se. The interpretation of the coefficient $\boldsymbol{B}^\circ_h$, on the other side, remains unchanged.

ii) Multiple Models Comprising an Intercept and Multiple Auxiliary Variables ($Q > 2$)
This section aims at extending the findings about the decomposition of B°_h in (5.21) to multiple models comprising an intercept and multiple auxiliary variables ($Q > 2$). Once again, we

decompose the multiple coefficient $\boldsymbol{B}_{\boldsymbol{h}}^{\circ}$ into a coefficient $\boldsymbol{B_h}$ from a reduced household-level model without an intercept and a remaining term capturing the effect of the intercept on $\boldsymbol{B}_{\boldsymbol{h}}^{\circ}$ that was disregarded previously. To quantify the remaining term, we translate the findings from paragraph i) about the mediation model to multiple models. The proof of the following lemma is kept short as it is based on the same arguments as in the simple model.

Lemma 1. ***Decomposition of the Integrated Household-Level Coefficient***
The integrated household-level coefficient $\boldsymbol{B}_{\boldsymbol{h}}^{\circ} = (B_{x_0}^{\circ}, \boldsymbol{B}_{\boldsymbol{x}}^{\circ})^T$ resulting from model $y_g = B_{x_0}^{\circ} x_{g0} + \boldsymbol{B}_{\boldsymbol{x}}^{\circ T} \boldsymbol{x_g} + r_g^{B^{\circ}}$ can be decomposed into

$$\boldsymbol{B}_{\boldsymbol{h}}^{\circ} = \begin{pmatrix} B_{x_0}^{\circ} \\ \boldsymbol{B}_{\boldsymbol{x}}^{\circ} \end{pmatrix} = \begin{pmatrix} B_{x_0}^{\circ} \\ \boldsymbol{B_h} - B_{x_0}^{\circ} \boldsymbol{F_x} \end{pmatrix}, \tag{5.22}$$

where $\boldsymbol{B_h}$ is the coefficient vector from the reduced household model: $y_g = \boldsymbol{B_h}^T \boldsymbol{x_g} + r_g^{B_h}$. Coefficient vector $\boldsymbol{F_x}$ results from the auxiliary model, $x_{g0} = \boldsymbol{F_x}^T \boldsymbol{x_g} + r_g^{F_x}$, which regresses the intercept on the remaining auxiliaries.

Proof. According to the mediation model introduced in paragraph i) the multiple total effect $\boldsymbol{B_h}$ can be split into the direct effect, $\boldsymbol{B}_{\boldsymbol{x}}^{\circ}$, of $\boldsymbol{x_g}$ on y_g controlled for x_{g0} and the indirect effect, $B_{x0}^{\circ} \boldsymbol{F_x}$, which constitutes the effects of the auxiliaries $\boldsymbol{x_g}$ on y_g via the intercept x_{g0}. Resolving yields

$$\underset{(Q\times 1)}{\boldsymbol{B}_{\boldsymbol{x}}^{\circ}} = \underset{(Q\times 1)}{\boldsymbol{B_h}} - \underset{(1\times 1)}{B_{x0}^{\circ}} \underset{(Q\times 1)}{\boldsymbol{F_x}} . \tag{5.23}$$

□

Lemma 1 provides the first part for the solution of Problem 1, as it enables us to separate the effect of the intercept from the integrated household-level coefficient $\boldsymbol{B}_{\boldsymbol{h}}^{\circ}$. In econometrics the technique of splitting a total effect into a direct and an indirect effect is sometimes called orthogonalization, for example, in Seber (1977).

5.2.1.2 Decomposition of the Integrated Household-Level Residuals

In the previous paragraph, we clarify the decomposition of the integrated coefficient $\boldsymbol{B}_{\boldsymbol{h}}^{\circ}$. In this paragraph, we translate the decomposition to the residuals, since we initially aim to separate the effect of the intercept from the variance of an integrated GREG estimator. The following lemma shows that the total sum of the decomposed residuals resulting from the separated regressions equals the residual from the initial integrated model for each household $g \in U_h$.

Lemma 2. ***Decomposition of the Integrated Household-Level Residuals***
The residuals from an integrated household-level GREG estimator $r_g^{B^\circ} = y_g - B^\circ_{x0} x_{g0} + \boldsymbol{B}_{\boldsymbol{x}}^{\circ T} \boldsymbol{x}_{\boldsymbol{g}}$ *can be decomposed into*

$$r_g^{B_h^\circ} = r_g^{B_h} + \tilde{r}_g^{B^\circ_{x0}} - \tilde{r}_g^{B^\circ_{x0} \cdot F_x} \tag{5.24}$$

where $r_g^{B_h} = y_g - {\boldsymbol{B}_{\boldsymbol{h}}}^T \boldsymbol{x}_{\boldsymbol{g}}$ *is the residual of the reduced household-level model. We denote* $\tilde{r}_g^{B^\circ_{x0}} = y_g - B^\circ_{x0} x_{g0}$ *and* $\tilde{r}_g^{B^\circ_{x0} \cdot F_x} = y_g - B^\circ_{x0} \cdot {\boldsymbol{F}_{\boldsymbol{x}}}^T \boldsymbol{x}_{\boldsymbol{g}}$ *as artificially constructed pseudo-residuals.*

Proof. Inserting the decomposition of the integrated household-level coefficient from Lemma 1 into the residuals $r_g^{B_h^\circ}$ for all $g \in U_h$ results in

$$\begin{aligned} r_g^{B_h^\circ} &= y_g - B^\circ_{x0} x_{g0} - \boldsymbol{B}_{\boldsymbol{x}}^{\circ T} \boldsymbol{x}_{\boldsymbol{g}} \\ &= y_g - B^\circ_{x0} x_{g0} - (\boldsymbol{B}_{\boldsymbol{h}} - B^\circ_{x0} \cdot {\boldsymbol{F}_{\boldsymbol{x}}}^T) \boldsymbol{x}_{\boldsymbol{g}}. \end{aligned}$$

Substitution of $\pm y_g$ yields

$$= \underbrace{(y_g - {\boldsymbol{B}_{\boldsymbol{h}}}^T \boldsymbol{x}_{\boldsymbol{g}})}_{r_g^{B_h}} + \underbrace{(y_g - B^\circ_{x0} x_{g0})}_{\tilde{r}_g^{B^\circ_{x0}}} - \underbrace{(y_g - B^\circ_{x0} \cdot {\boldsymbol{F}_{\boldsymbol{x}}}^T \boldsymbol{x}_{\boldsymbol{g}})}_{\tilde{r}_g^{B^\circ_{x0} \cdot F_x}}.$$

□

It should be noted that the equality in (5.24) is true for the entire M-vector of residuals. We define $\tilde{r}_g^{B^\circ_{x0}}$ and $\tilde{r}_g^{B^\circ_{x0} \cdot F_x}$ as **artificially constructed pseudo-residuals**, since they differ from residuals from ordinary regressions such as $r_g^{B_h} = y_g - {\boldsymbol{B}_{\boldsymbol{h}}}^T \boldsymbol{x}_{\boldsymbol{g}}$. Whereas the ordinary residual $r_g^{B_h}$ is given by the deviation between the observed value y_g and the hyperplane $\hat{y}_g = {\boldsymbol{B}_{\boldsymbol{h}}}^T \boldsymbol{x}_{\boldsymbol{g}}$ (or predicted values), pseudo-residuals are artificially constructed in a two-step procedure:

- In a first step, the coefficients B°_{x0} and $\boldsymbol{F}_{\boldsymbol{x}}$ are determined by the integrated household-level model $y_g = B^\circ_{x0} x_{g0} + \boldsymbol{B}_{\boldsymbol{x}}^{\circ T} \boldsymbol{x}_{\boldsymbol{g}} + r_g^{B_h^\circ}$ and the auxiliary model $x_{g0} = {\boldsymbol{F}_{\boldsymbol{x}}}^T \boldsymbol{x}_{\boldsymbol{g}} + r_g^{F_x}$, respectively.

- In a second step, based on these coefficients, we artificially construct the pseudo-residuals by $\tilde{r}_g^{B^\circ_{x0}} = y_g - B^\circ_{x0} x_{g0}$ and $\tilde{r}_g^{B^\circ_{x0} \cdot F_x} = y_g - B^\circ_{x0} \cdot {\boldsymbol{F}_{\boldsymbol{x}}}^T \boldsymbol{x}_{\boldsymbol{g}}$.

The original auxiliaries, constituting the coefficients B°_{x0} and $\boldsymbol{F}_{\boldsymbol{x}}$, do not fit to the auxiliaries used to construct the pseudo-residuals. The construction of pseudo-residuals permits us to exactly quantify the effect of the intercept on $r_g^{B_h^\circ}$ disregarded by $r_g^{B_h}$.

In the following, we derive some properties of residuals from regressions without an intercept and pseudo-residuals.

Properties of Residuals from Regressions without an Intercept. Omitting an intercept in a regression model significantly affects the properties of the residuals. To contrast the properties of the residuals with and without an intercept, we derive the **first normal equations** (cf. Greene, 2003, p. 243). We start with the integrated household-level model,

$$y_g = B^{\circ}_{x_0} x_{g0} - \boldsymbol{B^{\circ}_x}^T \boldsymbol{x_g} + r_g^{B^{\circ}_h},$$

as an ordinary model containing an intercept. Following the OLS theory, the first normal equation can be obtained by the first derivative of the sum of squared residuals with respect to B°_{x0}

$$\frac{\partial \sum_{g \in U_h} (r_g^{B^{\circ}_h})^2}{\partial B^{\circ}_{x_0}} = 2 \sum_{g \in U_h} \underbrace{(y_g - B^{\circ}_{x_0} x_{g0} - \boldsymbol{B^{\circ}_x}^T \boldsymbol{x_g})}_{r_g^{B^{\circ}_h}} (-x_{g0}).$$

Inserting $x_{g0} = 1$ for all $g \in U_h$ as intercept yields

$$= \sum_{g \in U_h} r_g^{B^{\circ}_h} \qquad\qquad \overset{!}{=} 0. \tag{5.25}$$

Accordingly, the first normal equation (5.25) states that the sum of squared residuals from a model containing an intercept is equal to zero.

In contrast, for models without an intercept the first normal equation is no longer valid. Instead, the first derivative of the sum of squared residuals of the reduced household model subject to B_h is obtained by

$$\frac{\partial \sum_{g \in U_h} (r_h^{B_h})^2}{\partial B_h} = 2 \sum_{g \in U_h} \underbrace{(y_g - \boldsymbol{B_h}^T \boldsymbol{x_g})}_{r_g^{B_h}} (-\boldsymbol{x_g}) \overset{!}{=} 0.$$

Consequently, since $x_{g1} \neq 1$ for all $g \in U_h$, the residuals of a model omitting the intercept no longer sum up to zero

$$\sum_{g \in U_h} r_g^{B_h} \neq 0. \tag{5.26}$$

Properties of Artificially Constructed Pseudo-Residuals. As aforementioned, artificially constructed pseudo-residuals do not result from ordinary regressions. Therefore, even if the initial coefficients $B^{\circ}_{x_0}$ and F_x, which are utilized to construct the pseudo-residuals, are determined by minimizing the sum of squared of residuals $r_g^{B^{\circ}_h}$ and $r_g^{F_x}$, the resulting constructed pseudo-residuals $\tilde{r}_g^{B^{\circ}_{x0}}$ and $\tilde{r}_g^{B^{\circ}_{x0} \cdot F_x}$ do not conform with the initial residuals $r_g^{B^{\circ}_h}$ and $r_g^{F_x}$. As a result, the normal equations of the initial residuals $r_g^{B^{\circ}_h}$ and $r_g^{F_x}$ are no longer valid. Thus, also the pseudo-residuals no longer sum up to zero

$$\sum_{g \in U_h} \tilde{r}_g^{B^{\circ}_{x0}} \neq 0 \quad \text{and} \quad \sum_{g \in U_h} \tilde{r}_g^{B^{\circ}_{x0} \cdot F_x} \neq 0. \tag{5.27}$$

The inequalities in (5.26) and (5.27) strongly influence the model unbiasedness of the estimated coefficients, because they rely on the assumption that the mean value of the residuals is in expectation equal to zero (cf. Wooldridge, 2013, p. 79). However, once more the only purpose of the residuals from the reduced household-level and the auxiliary model is the decomposition of the integrated residuals. Hence, we are not interested in the model properties of $r_g^{B_h}$, $\tilde{r}_g^{B^\circ_{x0}}$ and $\tilde{r}_g^{B^\circ_{x0} \cdot F_x}$ per se. Conversely, the model properties of the integrated household-level residuals $r_g^{B^\circ_h}$ remain unchanged. We refer the interested reader to an overview about model properties to econometric textbooks such as Greene (2003, Section 6.6) or Wooldridge (2013, Section 2.5).

We denote the residuals $r_g^{B_h}$, $\tilde{r}_g^{B^\circ_{x0}}$ and $\tilde{r}_g^{B^\circ_{x0} \cdot F_x}$ as **separating residuals** hereinafter, since they separate the effect of the intercept from $r_g^{B^\circ_h}$. The following remark states a powerful result that even if the single sum of the separating residuals is nonzero, their total sum, in turn, is equal to zero.

Remark 5. ***Total Sum of the Separating Residuals***
From the normal equation in (5.25) follows directly that

$$\bar{r}^{B_h} + \bar{\tilde{r}}^{B^\circ_{x0}} - \bar{\tilde{r}}^{B^\circ_{x0} \cdot F_x} = 0$$

with $\bar{r}^{B_h} = M^{-1} \sum_{g \in U_h} {r_g}^{B_h}$ *as the mean of the residuals from the reduced household-level model and* $\bar{\tilde{r}}^{B^\circ_{x0}}$ *as well as* $\bar{\tilde{r}}^{B^\circ_{x0} \cdot F_x}$ *in obvious notation.*

Remark 5 considerably simplifies the proof of the result in the following section.

5.2.1.3 Decomposition of the Sum of Squared Integrated Household-Level Residuals

Given the decomposition of the integrated residuals in Lemma 2, the following result provides the decomposition of the sum of squared integrated residuals as the last part of the solution for Problem 1. The decomposition is thusfar powerful, as even if the power of two is a non-linear transformation, the sum of the squared residuals of the original integrated household-level model equals the sum of squared residuals for the separated regressions, as it would be with a linear transformation. This fact crucially simplifies our calculations of the difference between the variances of an integrated household-level GREG estimator and a person-level GREG estimator in Section 5.2.2, because we can skip all mixed terms emerging when multiplying out the product in a binomial formula.

The following result permits us to compare the variances of a person-level and an integrated household-level GREG estimator, by comparing the variances of a person-level and a reduced household-level GREG estimator, both of the same dimension, and finally adding a constructed term which captures the effect of the intercept disregarded by the reduced household-level GREG estimator.

Result 8. ***Decomposition of the Sum of Squared Integrated Residuals***
The sum of squared residuals of an integrated household-level model can be decomposed into

$$\sum_{g\in U_h}\left(r_g^{B_h^\circ}\right)^2 = \sum_{g\in U_h}\left(r_g^{B_h}\right)^2 + \sum_{g\in U_h}\left(\tilde{r}_g^{B_{x0}^\circ}\right)^2 - \sum_{g\in U_h}\left(\tilde{r}_g^{B_{x0}^\circ\cdot F_x}\right)^2 \tag{5.28}$$

without any mixed terms emerging when solving binomial formulas. The residual of the reduced household-level model is determined by $r_g^{B_h} = y_g - \boldsymbol{B_h}^T\boldsymbol{x_g}$*. The pseudo-residuals are constructed by* $\tilde{r}_g^{B_{x0}^\circ} = y_g - B_{x0}^\circ x_{g0}$ *and* $\tilde{r}_g^{B_{x0}^\circ\cdot F_x} = y_g - B_{x0}^\circ \cdot \boldsymbol{F_x}^T\boldsymbol{x_g}$*.*

Proof. We start with inserting the decomposition of the residuals from Lemma 2 into the sum of squared residuals

$$\sum_{g\in U_h}\left(r_g^{B_h^\circ}\right)^2 = \sum_{g\in U_h}\left(r_g^{B_h} + \tilde{r}_g^{B_{x0}^\circ} - \tilde{r}_g^{B_{x0}^\circ\cdot F_x}\right)^2. \tag{5.29}$$

As inequalities (5.26) and (5.27) induce that the separating residuals $r_g^{B_h}$, $\tilde{r}_g^{B_{x0}^\circ}$ and $\tilde{r}_g^{B_{x0}^\circ\cdot F_x}$ have a nonzero mean, we actually have to subtract the corresponding mean values of the residuals in the variance formula. However, following Remark 5, the total sum of the mean values of the residuals equals zero, and thus we can neglect them. Solving the binominal formula in (5.29) yields

$$\begin{aligned}
&= \sum_{g\in U_h}\left(\left(r_g^{B_h}\right)^2 + \left(\tilde{r}_g^{B_{x0}^\circ}\right)^2 + \left(\tilde{r}_g^{B_{x0}^\circ\cdot F_x}\right)^2 + 2r_g^{B_h}\tilde{r}_g^{B_{x0}^\circ} - 2r_g^{B_h}\tilde{r}_g^{B_{x0}^\circ\cdot F_x} - 2\tilde{r}_g^{B_{x0}^\circ}\tilde{r}_g^{B_{x0}^\circ\cdot F_x}\right)\\
&= \sum_{g\in U_h}\left(r_g^{B_h}\right)^2 + \sum_{g\in U_h}\left(\tilde{r}_g^{B_{x0}^\circ}\right)^2 + \sum_{g\in U_h}\left(\tilde{r}_g^{B_{x0}^\circ\cdot F_x}\right)^2\\
&\qquad + 2\sum_{g\in U_h} r_g^{B_h}\tilde{r}_g^{B_{x0}^\circ} - 2\sum_{g\in U_h} r_g^{B_h}\tilde{r}_g^{B_{x0}^\circ\cdot F_x} - 2\sum_{g\in U_h}\tilde{r}_g^{B_{x0}^\circ}\tilde{r}_g^{B_{x0}^\circ\cdot F_x}.
\end{aligned}$$

To recreate the solution of the binominal expansion, we add $\pm(\tilde{r}_g^{B_{x0}\cdot F_x})^2$

$$\begin{aligned}
&= \sum_{g\in U_h}(r_g^{B_h})^2 + \sum_{g\in U_h}(\tilde{r}_g^{B_{x0}^\circ})^2 - \sum_{g\in U_h}(\tilde{r}_g^{B_{x0}^\circ\cdot F_x})^2\\
&\quad \underbrace{+2\sum_{g\in U_h}(\tilde{r}_g^{B_{x0}^\circ\cdot F_x})^2 + 2\sum_{g\in U_h} r_g^{B_h}\tilde{r}_g^{B_{x0}^\circ} - 2\sum_{g\in U_h} r_g^{B_h}\tilde{r}_g^{B_{x0}^\circ\cdot F_x} - 2\sum_{g\in U_h}\tilde{r}_g{}^{B_{x0}^\circ}\tilde{r}_g^{B_{x0}^\circ\cdot F_x}}_{\text{term } (I)}.
\end{aligned}$$

To prove equality (5.28), it remains to prove that term (I) equals zero. After some simple rearrangement of term (I) we get

$$\begin{aligned}
&\sum_{g\in U_h}(y_g - B^\circ_{x0}\boldsymbol{F_x}^T\boldsymbol{x_g})^2 + \sum_{g\in U_h}(y_g - \boldsymbol{B_h}^T\boldsymbol{x_g})(y_g - B^\circ_{x0}x_{g0}) \\
&\qquad - \sum_{g\in U_h}(y_g - \boldsymbol{B_h}^T\boldsymbol{x_g})(y_g - B^\circ_{x0}\boldsymbol{F_x}^T\boldsymbol{x_g}) - \sum_{g\in U_h}(y_g - B^\circ_{x0}x_{g0})(y_g - B^\circ_{x0}\boldsymbol{F_x}^T\boldsymbol{x_g}) \\
&= \sum_{g\in U_h} y_g^2 - 2\sum_{g\in U_h} y_g B^\circ_{x0}\boldsymbol{F_x}^T\boldsymbol{x_g} + \sum_{g\in U_h} {B^\circ_{x0}}^2\boldsymbol{F_x}^T\boldsymbol{x_g}\boldsymbol{F_x}^T\boldsymbol{x_g} \\
&\qquad + \sum_{g\in U_I} y_g^2 - \sum_{g\in U_h} y_g\boldsymbol{B_h}^T\boldsymbol{x_g} - \sum_{g\in U_h} y_g B^\circ_{x0}x_{g0} + \sum_{g\in U_h} B^\circ_{x0}x_{g0}\boldsymbol{B_h}^T\boldsymbol{x_g} \\
&\qquad - \sum_{g\in U_h} y_g^2 + \sum_{g\in U_h} y_g B^\circ_{x0}\boldsymbol{F_x}^T\boldsymbol{x_g} + \sum_{g\in U_h} y_g\boldsymbol{B_h}^T\boldsymbol{x_g} - \sum_{g\in U_h} B^\circ_{x_{g0}}\boldsymbol{F_x}^T\boldsymbol{x_g}\boldsymbol{B_h}^T\boldsymbol{x_g} \\
&\qquad - \sum_{g\in U_h} y_g^2 + \sum_{g\in U_h} y_g B^\circ_{x0}\boldsymbol{F_x}^T\boldsymbol{x_g} + \sum_{g\in U_h} B^\circ_{x0}x_{g0}y_g - \sum_{g\in U_h} {B^\circ_{x0}}^2 x_{g0}\boldsymbol{F_x}^T\boldsymbol{x_g}.
\end{aligned}$$

Substituting the fact that x_{g0} is constant and thus $x_{g0}\sum_{g\in U_h}\boldsymbol{x_g} = \sum_{g\in U_h}\boldsymbol{x_g}$, we obtain

$$= \sum_{g\in U_h} {B^\circ_{x0}}^2\boldsymbol{F_x}^T\boldsymbol{x_g}\boldsymbol{F_x}^T\boldsymbol{x_g} + \sum_{g\in U_h} B^\circ_{x0}\boldsymbol{B_h}^T\boldsymbol{x_g} - \sum_{g\in U_h} {B^\circ_{x0}}^2\boldsymbol{F_x}^T\boldsymbol{x_g} - \sum_{g\in U_h} B^\circ_{x0}\boldsymbol{F_x}^T\boldsymbol{x_g}\boldsymbol{B_h}^T\boldsymbol{x_g}. \tag{5.30}$$

Exploiting the following relation

$$\begin{aligned}
\sum_{g\in U_h}\boldsymbol{F_x}\boldsymbol{x_g}^T\boldsymbol{x_g} &= \Big(\sum_{g\in U_h}\boldsymbol{x_g}\boldsymbol{x_g}^T\Big)^{-1}\sum_{g\in U_h}\boldsymbol{x_g}x_{g0}\sum_{g\in U_h}\boldsymbol{x_g}^T\boldsymbol{x_g} \\
&= \sum_{g\in U_h}\boldsymbol{x_g}
\end{aligned}$$

it becomes evident that (5.30) simplifies to

$$= \sum_{g\in U_h} {B^\circ_{x0}}^2\boldsymbol{F_x}^T\boldsymbol{x_g} + \sum_{g\in U_h} B^\circ_{x0}\boldsymbol{B_h}^T\boldsymbol{x_g} - \sum_{g\in U_h} {B^\circ_{x0}}^2\boldsymbol{F_x}^T\boldsymbol{x_g} - \sum_{g\in U_h} B^\circ_{x0}\boldsymbol{B_h}^T\boldsymbol{x_g}.$$

Therefore, term (I) is equal to zero, and (5.28) is proven. □

To conclude, Result 8 provides the final solution for Problem 1 of separating the effect of the intercept from the integrated household-level variance. It not only permits a correct comparison of the variances of an integrated household-level and a person-level GREG estimator, but it also allows in general a comparison of variances of GREG estimators of different dimensions. By artificially constructing pseudo-residuals, it is possible to exactly quantify the difference in the variances and attribute it to the effects of the additional auxiliary variable(s). In Section 5.3, we propose a further application of the decomposition derived in Result 8 in order to predict the difference between two coefficients of determination when adding or omitting explanatory variables.

5.2.2 Inserting the Decomposition of the Sum of Squared Residuals into the Efficiency Comparison

To continue with our initial aim to provide a correct comparison of the variances of the person-level and integrated household-level GREG estimators, we insert Result 8 into the objective function (5.13) from the beginning of the section

$$
\begin{aligned}
&\left(\mathrm{V}(\hat{T}_y^{\text{GREG}}) - \mathrm{V}(\hat{T}_y^{\text{INT}})\right) \Big/ \frac{M^2}{m}\left(1-\frac{m}{M}\right)(M-1)^{-1} \\
&\quad = \sum_{g\in U_h}\left(r_g^{B_p}\right)^2 - \sum_{g\in U_h}\left(r_g^{B_h^\circ}\right)^2 \\
&\quad = \sum_{g\in U_h}\left(r_g^{B_p}\right)^2 - \sum_{g\in U_h}\left(r_g^{B_h}\right)^2 - \sum_{g\in U_h}\left(\tilde{r}_g^{B_{x0}^\circ}\right)^2 + \sum_{g\in U_h}\left(\tilde{r}_g^{B_{x0}^\circ\cdot F_x}\right)^2 \\
&\quad = \underbrace{\sum_{g\in U_h}\left(y_g - \boldsymbol{B_p}^T\boldsymbol{x_g}\right)^2 - \sum_{g\in U_h}\left(y_g - \boldsymbol{B_h}^T\boldsymbol{x_g}\right)^2}_{I} \\
&\qquad\qquad - \underbrace{\left(\sum_{g\in U_h}(y_g - B_{x0}^\circ x_{g0})^2 - \sum_{g\in U_h}(y_g - B_{x0}^\circ\cdot\boldsymbol{F_x}^T\boldsymbol{x_g})^2\right)}_{\text{Effect of the intercept}}. \qquad (5.31)
\end{aligned}
$$

Thus, we have successfully quantified the remaining term capturing the effect of the intercept disregarded by the reduced household-level model. The effects related to the intercept, excluded from the integrated GREG estimator, are captured by the pseudo-residuals $\tilde{r}_g^{B_{x0}^\circ}$ and $\tilde{r}_g^{B_{x0}^\circ\cdot F_x}$. Once the effect of the intercept is separated, the person-level model and the reduced household-model in *I* are comparable in dimension, because the auxiliaries of the person-level model sum up to the auxiliaries of a reduced household-level model. Now, to further simplify term *I* in (5.31) we seek in the following section a relationship between $\boldsymbol{B_p}$, the coefficient of a person model, and $\boldsymbol{B_h}$, the coefficient of the reduced household model.

Remark 6. ***Alternative Proceeding for the Efficiency Comparison***
Instead of separating the effect of the intercept from the integrated household-level model, we could alternatively add an additional auxiliary, N_g^{-1}, to the person-level model. This temporarily added auxiliary variable sums up per household to the household-level intercept $x_{g1} = 1$. Therefore, we are able to compare an augmented person-level model of dimension $(Q+1)$ with the integrated household-level model of the same dimension. Finally, we have to subtract the effect of the temporarily added auxiliary variable from the variance of the augmented person-level model. However, we refrain from the alternative proceeding due to aforementioned advantage that the intercept in a household model is interpretable.

5.2.3 Relationship between the Coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$

After successfully separating the effect of the intercept from the integrated household-level model, the person-level and the reduced household models in Equation (5.13) are of the same dimension. To assess the difference between both variances, we seek a functional relationship between the coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$, since we doubt the appropriateness of the interpretation of $\boldsymbol{B_h} - \boldsymbol{B_p} = \boldsymbol{B_c}$ given by Steel and Clark (2007) (as argued in detail in Section 5.1.2.2). This issue is formalized by the following problem.

Problem 2. ***The Functional Relationship between B_p and B_h***
We are interested in a relationship between the coefficient of a person-level model $\boldsymbol{B_p}$, resulting from $y_i = \boldsymbol{B_p}^T \boldsymbol{x_i} + r_i^{B_p}$, and the coefficient of a reduced household-level model $\boldsymbol{B_h}$, resulting from $y_g = \boldsymbol{B_h}^T \boldsymbol{x_g} + r_g^{B_h}$. In other words, we aim to write $\boldsymbol{B_p}$ as function of $\boldsymbol{B_h}$, or vice versa.

Table 5.1 summarizes the models generating the coefficients under consideration $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$. In the following, this table is continued to outline our proceeding.

Table 5.1: Models under consideration to derive a relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ I

Person-level model	**Reduced household-level model**
$y_i = \boldsymbol{B_p}^T \boldsymbol{x_i} + r_i^{B_p}$ with $\boldsymbol{B_p} = (B_{p_1}, B_{p_2})^T$	$y_g = \boldsymbol{B_h}^T \boldsymbol{x_g} + r_g^{B_h}$ with $\boldsymbol{B_h} = (B_{h_1}, B_{h_2})^T$

The idea to solve Problem 2 is to relate the coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ by an overlap model that simultaneously contains the auxiliaries of both coefficients (see Table 5.1). Then, we decompose $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ into the same coefficients obtained from such an overlap model. For the decomposition of $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$, we apply the mediation model, introduced in Section 5.2.1.1. In the end of this section, we can write one coefficient as function of the other coefficient.

For a better understanding, we visualize our proceeding to derive a functional relationship with diagrams. Therefore, in Section 5.2.3.1, we initially focus on simple models comprising an intercept and one single auxiliary variable. Subsequently, we extend our findings to multiple models with $Q > 2$ auxiliary variables (Section 5.2.3.2).

5.2.3.1 Simple Models Comprising an Intercept and a Single Auxiliary Variable ($Q = 2$)

We start with the case of simple models comprising an intercept x_{i1} and one auxiliary variable x_{i2}. As done before, we define $\boldsymbol{x_i} = (x_{i1}, x_{i2})^T = (1, x_{i2})^T$ as auxiliary vector of a person-level GREG estimator. It contains the person-level intercept $x_{i1} = 1$. The auxiliaries of a reduced household-level model is determined by $\boldsymbol{x_g} = (x_{g1}, x_{g2})^T = (N_g, x_{g2})^T$. Thus, it omits an household-level intercept, $x_{g0} = 1$.

The fact that the person-level and the reduced household-level GREG estimators are constructed at different estimation levels (see Table 5.1) hampers finding a relationship between its coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$. To handle this obstacle, we exploit our finding, shown in Section 3.1.4 in

(3.8). According to this, the integrated household-level coefficient $\boldsymbol{B_h}$ can be calculated either by OLS or GLS. Translating this finding to the reduced household-level coefficient $\boldsymbol{B_h}$, we obtain that

$$\boldsymbol{B_h} = \Big(\sum_{g\in U_h} \boldsymbol{x_g}\boldsymbol{x_g}^T\Big)^{-1} \sum_{g\in U_h} \boldsymbol{x_g} y_g, \tag{5.32}$$

arising from the household-level model $y_g = \boldsymbol{B_h}^T\boldsymbol{x_g} + r_g^{B_h}$ using OLS, coincides with the coefficient derived at the person level,

$$\begin{aligned}
&= \Big(\sum_{g\in U_h} N_g\boldsymbol{\bar{x}_i} N_g\boldsymbol{\bar{x}_i}^T\Big)^{-1} \sum_{g\in U_h}\sum_{i\in U_g} N_g\boldsymbol{\bar{x}_i} y_g \\
&= \Big(\sum_{g\in U_h}\sum_{i\in U_g} N_g\boldsymbol{\bar{x}_i} N_g\boldsymbol{\bar{x}_i}^T N_g^{-1}\Big)^{-1} \sum_{g\in U_h}\sum_{i\in U_g} N_g\boldsymbol{\bar{x}_i} y_i \\
&= \Big(\sum_{i\in U_p} N_g\boldsymbol{\bar{x}_i}\boldsymbol{\bar{x}_i}^T\Big)^{-1} \sum_{i\in U_p} N_g\boldsymbol{\bar{x}_i} y_i,
\end{aligned} \tag{5.33}$$

arising from the model $y_i = \boldsymbol{B_h}^T\boldsymbol{\bar{x}_i} + r_i^{B_h}$ using GLS. To trace the rearrangement note that $\boldsymbol{x_g} = N_g\boldsymbol{\bar{x}_i}$ and $\sum_{i\in U_p} = \sum_{g\in U_h}\sum_{i\in U_g}$. The person-level auxiliary vector of an integrated GREG estimator is determined by $\boldsymbol{\bar{x}_i} = (\bar{x}_{i1}, \bar{x}_{i2})^T = (1, \bar{x}_{i2})^T$. In consequence, even if the initial calculation levels differ, we can use two person models based on $\boldsymbol{x_i}$ or $\boldsymbol{\bar{x}_i}$ in order to derive a relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$. The reduced model at the person level is called the **reduced person-level model**. Table 5.2 summarizes this result. To keep notation simple, we refrain from denoting the computational level of $\boldsymbol{B_h}$ in (5.32) or (5.33) by an extra index. An extra index would falsely imply that both coefficients are numerically different. Note that the left-hand side of Table 5.2 remains unchanged.

Table 5.2: Models under consideration to derive a relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ II

Person-level model	**Reduced household-level model**
$y_i = \boldsymbol{B_p}^T\boldsymbol{x_i} + r_i^{B_p}$	$y_g = \boldsymbol{B_h}^T\boldsymbol{x_g} + r_g^{B_h}$
The relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ can alternatively be derived by	
Person-level model	**Reduced person-level model**
$y_i = \boldsymbol{B_p}^T\boldsymbol{x_i} + r_i^{B_p}$	$y_i = \boldsymbol{B_h}^T\boldsymbol{\bar{x}_i} + r_i^{B_h}$

It is important to note that the reduced model on the lower right-hand side of Table 5.2 indeed contains an person-level intercept. The reason for this is that in order to obtain the reduced model at the person level on the lower right-hand side, we have to inflate the auxiliary vector $\boldsymbol{x_g}$ from the model on the upper right side by N_g^{-1}. Hence, it follows that $N_g^{-1}\boldsymbol{x_g} =$

$N_g^{-1}(x_{g1}, x_{g2})^T = N_g^{-1}(N_g, x_{g2})^T$ results in $\bar{\boldsymbol{x}}_i = (\bar{x}_{i1}, \bar{x}_{i2})^T = (1, \bar{x}_{i2})^T$. According to this, $\bar{\boldsymbol{x}}_i$ contains a person-level intercept $x_{i1} = 1$ as counterpart to the household-level auxiliary $\bar{x}_{g1} = N_g$. However, the term *reduced* in the model on the upper right hand-side of Table 5.2 refers to the household-level intercept denoted as $x_{g0} = 1$. Thus, once more it is important to differentiate between person- and household-level intercepts.

Now, the idea is to derive a relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ by utilizing a model that simultaneously contains the auxiliaries $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_i$ and thus determines an overlap between both models at hand. In the context of multilevel studies, models are known that include the mean values of the auxiliaries as additional explanatory variables to account for possible correlations between the individual and the group level (cf. Wooldridge, 2013, p. 479; Gelman, 2006, p. 434). However, the classical multilevel model assumes a random effect structure at the individual level, caused by the group to which the individual belongs (cf. Snijders, 2011; Stryhn et al., 2006). Nevertheless, we are interested only in relating $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ by modeling the overlap between the person-level and the integrated model. We are not interested in explaining a variable of interest by group effects. Thus, we appropriated the idea of incorporating $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_i$ as auxiliaries, but we do not assume an underlying random effect structure, which is the core idea of multilevel model. Hence, we denote such models in the following as **overlap models**. The overlap model in the one-dimensional case can be expressed by

$$y_i = D_{x_1} x_{i1} + D_{x_2} x_{i2} + D_{\bar{x}_2} \bar{x}_{i2} + r_i^D. \tag{5.34}$$

It comprises x_{i2} and its corresponding household mean value $\bar{x}_{i2}$, both referring to the same variable. Then, based on the overlap model, we decompose $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$, depending on either $\boldsymbol{x}_i$ or $\bar{\boldsymbol{x}}_i$, into the same coefficients D_{x_1}, D_{x_2} and $D_{\bar{x}_2}$. For the decomposition, we apply two mediation models extensively discussed in Section 5.2.1.1. Finally, solving $\boldsymbol{B_h}$ for D_{x_2} and inserting into $\boldsymbol{B_p}$ permits one to write $\boldsymbol{B_p}$ as function of $\boldsymbol{B_h}$. The detailed calculations are given in the proof of Result 9.

At this point, it becomes obvious why we transform the reduced household-level model to person-level (see Table 5.2) rather than reverse-transform the person-level model to the household level. The reason for this choice is that only at the person level, can we relate the coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ by an overlap model. Instead, when transforming the reverse direction, and thus deriving the relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ at the household level, both auxiliaries $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_i$ would sum up per household to the same vector $\boldsymbol{x_g} = \sum_{i \in U_g} \boldsymbol{x}_i = \sum_{i \in U_g} \bar{\boldsymbol{x}}_i$. Therefore, at the household level we were not able to model the overlap. For this reason, we decide to derive the relationship at the person level. Notwithstanding, even when the relationship is derived at the person level, it is also valid at the household level.

It is important to differentiate the proceeding in this section from the proceeding in the previous section. In Section 5.2.1, we separate the effect of the intercept from the variance of an integrated household-level GREG estimator. Since we deal with household surveys and since the variance under cluster sampling refers to the aggregates of the variables, we derived the decomposition at the household level. In this section, on the other hand, we are interested in deriving a relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ by exploiting an overlap model at the person level.

Unfortunately solving B_p and D_{x_2} from the overlap model (5.34) is not straightforward, because the arrays of the decomposed coefficients differ, as we will see later in this section. To handle this obstacle, we derive Lemma 3. It describes the form of the coefficient resulting from regressing the original auxiliary x_{i2} on its constructed household mean values $\bar{x}_{i2}$, whereas x_{i2} and $\bar{x}_{i2}$ refer to the same variable.

Lemma 3. ***The Form of the Coefficient $B_{\bar{x}}$ from Regressing an Original Auxiliary on Its Constructed Mean Values***
The coefficient vector $\boldsymbol{B}_{\bar{\boldsymbol{x}}}$ obtained from the model $x_{i2} = B_{\bar{x}_1}\bar{x}_{i1} + B_{\bar{x}_2}\bar{x}_{i2} + r_i^{B_{\bar{x}}}$ is given by

$$\boldsymbol{B}_{\bar{\boldsymbol{x}}} = \begin{pmatrix} B_{\bar{x}_1} \\ B_{\bar{x}_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{5.35}$$

Note that $\bar{x}_{i1} = 1$ is the intercept in the model.

Proof. The coefficients of a bivariate regression $x_{i2} = B_{\bar{x}_1}\bar{x}_{i1} + B_{\bar{x}_2}\bar{x}_{i2} + r_i^{B_{\bar{x}}}$ are defined as

$$B_{\bar{x}_2} = \frac{\text{Cov}(\boldsymbol{y}, \boldsymbol{x_2})}{\text{Var}(\boldsymbol{x_2})} \tag{5.36}$$

$$B_{\bar{x}_1} = \bar{y} - B_{\bar{x}_2}\bar{x}_2 \tag{5.37}$$

with $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ and $\boldsymbol{x_2} = (x_{21}, \ldots, x_{2N})^T$ (cf. von Auer, 2007, p. 58). $\bar{y} = N^{-1}\sum_{i\in U_p} y_i$ and $\bar{x}_2 = N^{-1}\sum_{i\in U_p}\bar{x}_{i2}$ are denoted as mean values. The coefficient in (5.36) describes the slope parameter, the coefficient in (5.37) defines the intercept.

To validate (5.35), we start to show that $B_{\bar{x}2} = 1$. Hence, we have to prove that the numerator and denominator of the slope parameter,

$$B_{\bar{x}_2} = \frac{\sum_{i\in U_p}(x_{i2} - \bar{x}_2)(\bar{x}_{i2} - \bar{x}_2)}{\sum_{i\in U_p}(\bar{x}_{i2} - \bar{x}_2)^2}, \tag{5.38}$$

are equal[4]. Since the totals of x_{i2} and $\bar{x}_{i2}$ are equal, we define $\bar{x}_2$ as mean value of both auxiliaries.

Preliminarily, we verify two equalities

$$\begin{aligned} \sum_{i\in U_p} x_{i2}\bar{x}_{i2} &= \sum_{i\in U_p}\bar{x}_{i2}\sum_{i\in U_g}\frac{x_{i2}}{N_g} = \sum_{i\in U_p}\bar{x}_{i2}^2 \\ \sum_{i\in U_p}\bar{x}_2 x_{i2} &= \sum_{i\in U_p}\bar{x}_2\sum_{i\in U_g}\frac{x_{i2}}{N_g} = \sum_{i\in U_p}\bar{x}_{i2}\bar{x}_2. \end{aligned} \tag{5.39}$$

Note that the totals of x_{i2} and $\bar{x}_{i2}$ are equal. Given these equalities it is easy to verify that the numerator and denominator in (5.38) are equal such that

$$\begin{aligned} \sum_{i\in U_p}(x_{i2} - \bar{x}_2)(\bar{x}_{i2} - \bar{x}_2) &= \sum_{i\in U_p} x_{i2}\bar{x}_{i2} - \sum_{i\in U_p}\bar{x}_2 x_{i2} - \sum_{i\in U_p}\bar{x}_{i2}\bar{x}_2 + \sum_{i\in U_p}\bar{x}_2^2 \\ &= \sum_{i\in U_p}\bar{x}_{i2}^2 - 2\sum_{i\in U_p}\bar{x}_{i2}\bar{x}_2 + \sum_{i\in U_p}\bar{x}_2^2 \\ &= \sum_{i\in U_p}(\bar{x}_{i2} - \bar{x}_2)^2. \end{aligned}$$

[4] The term $1/(N-1)$ emerges in both the numerator and the denominator and thus is canceled out.

Thus, $B_{\bar{x}_2} = 1$. It remains to be shown that $B_{\bar{x}_1} = 0$. Inserting $B_{\bar{x}_2} = 1$ into (5.37), we obtain

$$\begin{aligned} B_{\bar{x}2} &= \frac{1}{N} \sum_{i \in U_p} x_{i2} - B_{\bar{x}2} \frac{1}{N} \sum_{i \in U_p} \bar{x}_{i2} \\ &= \bar{x}_2 - \bar{x}_2 \\ &= 0. \end{aligned}$$

Therefore, (5.35) is proven. □

The result of the form of $\boldsymbol{B}_{\bar{\boldsymbol{x}}}$ in Lemma 3 is not surprising. For explanation, consider x_i denotes the variable `sex` with value 1 if person i is a woman, and with value 2 if person i is a man. The original value for `sex`, x_i, is regressed on the household mean value for `sex`, $\bar{x}_i$. At first, in case of $\bar{x}_i = 0$ for all $i \in U_g$ all household members are male, which implies that the regression line runs through the origin. This is equivalent with $B_{x1} = 0$. Secondly, a slope of $B_{x2} = 1$ induces that a change from 0 to 1 of $\bar{x}_i$ leads to a change from 0 to 1 of the x_i for all $i \in U_g$. This is because a change from 0 to 1 of the household mean value for `sex` implies that all household members are either women or men. It should be noted that the result of Lemma 3 is valid only for regressing x_i on $\bar{x}_i$, but not for the reverse case of regressing $\bar{x}_i$ on x_i.

Now, we continue with deducing a functional relationship between the coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$. The idea is to exploit the fact that an overlap model comprises both auxiliaries $\boldsymbol{x_i}$ and $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$. We then decompose $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ into the same coefficients obtained from the overlap model. For the decomposition of the coefficients, we apply two mediation models introduced in Section 5.2.1.1 and interpret $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ as direct effects. The detailed explanation of the proceeding is given within the proof. For a better understanding, we turn away from a classical proof and include graphs and tables to underpin our argumentation.

Result 9. ***The Functional Relationship between $\mathbf{B}_h$ and $\mathbf{B}_p$ for Simple Models***
The coefficient $\mathbf{B}_p$, resulting from the person-level model $y_i = \boldsymbol{B_p}^T \boldsymbol{x_i} + r_i^{B_p}$, can be expressed as

$$\boldsymbol{B_p} = \boldsymbol{B_h} + D_{\bar{x}_2}(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}}),$$

where $\boldsymbol{B_h}$ results from the reduced person-level model $y_i = \boldsymbol{B_h}^T \bar{\boldsymbol{x}}_{\boldsymbol{i}} + r_i^{B_h}$. $D_{\bar{x}_2}$ arises from the overlap model $y_i = D_{x_1} x_{i1} + D_{x_2} x_{i2} + D_{\bar{x}_2} \bar{x}_{i2} + r_i^D$ and describes the overlap between the person- and household-level auxiliaries. $\boldsymbol{B_x}$ and $\boldsymbol{B_{\bar{x}}}$ are obtained from the auxiliary models $\bar{x}_i = \boldsymbol{B_x}^T \boldsymbol{x_i} + r_i^{B_{\bar{x}}}$ and $x_i = \boldsymbol{B_{\bar{x}}}^T \bar{\boldsymbol{x}}_{\boldsymbol{i}} + r_i^{B_x}$, respectively.

Proof. We start with decomposing the coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ with respect to the following overlap model

$$y_i = D_{x_1} x_{i1} + D_{x_2} x_{i2} + D_{\bar{x}_2} \bar{x}_{i2} + r_i^D. \tag{5.40}$$

For this purpose, we apply two mediation models, as introduced in Section 5.2.1: one for the decomposition of $\boldsymbol{B_p}$ (see Figure 5.10) and one for the decomposition of $\boldsymbol{B_h}$ (see Figure 5.11).

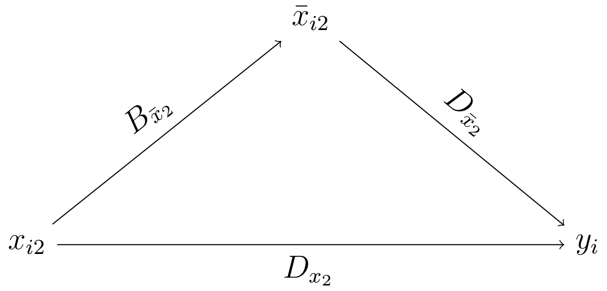

Figure 5.10: Mediation model applied to the auxiliary model I

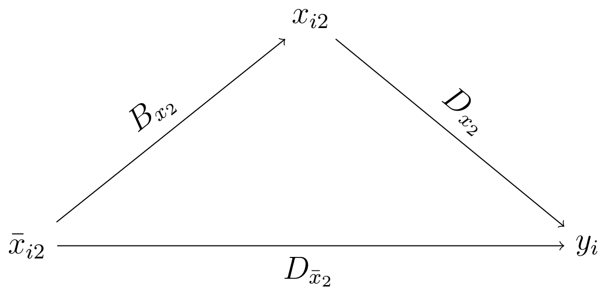

Figure 5.11: Mediation model applied to the auxiliary model II

The difference between the Figures 5.10 and 5.11 is the direction of the auxiliary regression, meaning whether $\bar{x}_i$ is regressed on x_i (results in B_x) or vice versa (results in $B_{\bar{x}}$). Nevertheless, the coefficients D_{x_2} and $D_{\bar{x}_2}$ refer in both figures to the same overlap model (5.40).

In order to decompose $\boldsymbol{B_p}$, we interpret $\bar{x}_{i2}$ as mediator variable. Then, in accordance with Figure 5.10, we split the total effect of $\boldsymbol{B_p}$ obtained from regressing y_i on x_{i2}, into the direct effect of x_{i2} on y_i, controlling for the mediator variable $\bar{x}_{i2}$ and the indirect effect of x_{i2} via $\bar{x}_{i2}$. To quantify the indirect effect, we specify the auxiliary model $\bar{x}_i = \boldsymbol{B_x}^T \boldsymbol{x_i} + r_i^{B_{\bar{x}}}$ with $\boldsymbol{B_x} = (B_{x_1}, B_{x_2})^T$. Hence, the total effect $\boldsymbol{B_p}$ can be decomposed into

$$\boldsymbol{B_p} = \begin{pmatrix} B_{p_1} \\ B_{p_2} \end{pmatrix} = \begin{pmatrix} D_{x_1} \\ D_{x_2} \end{pmatrix} + D_{\bar{x}_2} \cdot \begin{pmatrix} B_{x_1} \\ B_{x_2} \end{pmatrix}.$$

We analogously proceed with the decomposition of the household-level coefficient $\boldsymbol{B_h}$ according to Figure 5.11. Table 5.3 summarizes the auxiliary models and resulting decomposed coefficients from a person-level model and a reduced person-level model.

Table 5.3: Models under consideration to derive a relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ III

<table>
<tr><th>Person-level model</th><th>Reduced person-level model</th></tr>
<tr><td>$y_i = \boldsymbol{B_p}^T \boldsymbol{x_i} + r_i^{B_p}$
with $\boldsymbol{B_p} = (B_{p_1}, B_{p_2})^T$</td><td>$y_i = \boldsymbol{B_h}^T \bar{\boldsymbol{x}}_{\boldsymbol{i}} + r_i^{B_h}$
with $\boldsymbol{B_h} = (B_{h_1}, B_{h_2})^T$</td></tr>
<tr><td colspan="2">The auxiliary models for the decomposition are given by</td></tr>
<tr><td>$\bar{x}_i = \boldsymbol{B_x}^T \boldsymbol{x_i} + r_i^{B_{\bar{x}}}$
with $\boldsymbol{B_x} = (B_{x_1}, B_{x_2})^T$</td><td>$x_i = \boldsymbol{B_{\bar{x}}}^T \bar{\boldsymbol{x}}_{\boldsymbol{i}} + r_i^{B_x}$
with $\boldsymbol{B_{\bar{x}}} = (B_{\bar{x}_1}, B_{\bar{x}_2})^T$</td></tr>
<tr><td colspan="2">The coefficients are decomposed into</td></tr>
<tr><td>$\boldsymbol{B_p} = \begin{pmatrix} B_{p_1} \\ B_{p_2} \end{pmatrix} = \begin{pmatrix} D_{x_1} \\ D_{x_2} \end{pmatrix} + D_{\bar{x}_2} \cdot \begin{pmatrix} B_{x_1} \\ B_{x_2} \end{pmatrix}$</td><td>$\boldsymbol{B_h} = \begin{pmatrix} B_{h_1} \\ B_{h_2} \end{pmatrix} = \begin{pmatrix} D_{x_1} \\ D_{\bar{x}_2} \end{pmatrix} + D_{x_2} \cdot \begin{pmatrix} B_{\bar{x}_1} \\ B_{\bar{x}_2} \end{pmatrix}$</td></tr>
</table>

As result, we can write $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ as functions of D_{x_1}, D_{x_2} as well as $D_{\bar{x}_2}$ among others. Unfortunately, the positions of D_{x_2} and $D_{\bar{x}_2}$ in the arrays differ between the right- and the left-hand side in Table 5.3. To solve this issue, we exploit the form of $\boldsymbol{B_{\bar{x}}}$ derived in Lemma 3. According to this, we can resort the array of the household coefficient, because

$$\begin{pmatrix} B_{h_1} \\ B_{h_2} \end{pmatrix} = \begin{pmatrix} D_{x_1} \\ D_{\bar{x}_2} \end{pmatrix} + D_{x_2} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

is equivalent to

$$= \begin{pmatrix} D_{x_1} \\ D_{x_2} \end{pmatrix} + D_{\bar{x}_2} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} .$$

The positions of D_{x_1} and $D_{\bar{x}_2}$ in the first and second line are swapped. Consequently, coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ are both functions of the same

$$\boldsymbol{D_x} = \begin{pmatrix} D_{x_1} \\ D_{x_2} \end{pmatrix} \quad \text{and} \quad D_{\bar{x}_2}. \tag{5.41}$$

Table 5.4 outlines the resorting of $\boldsymbol{B_h}$ and continues the Table 5.3.

Table 5.4: Models under consideration to derive a relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ IV

Person-level model	**Reduced person-level model**
$\boldsymbol{B_p} = \begin{pmatrix} B_{p_1} \\ B_{p_2} \end{pmatrix} = \begin{pmatrix} D_{x_1} \\ D_{x_2} \end{pmatrix} + D_{\bar{x}_2} \cdot \begin{pmatrix} B_{x_1} \\ B_{x_2} \end{pmatrix}$	$\boldsymbol{B_h} = \begin{pmatrix} B_{h_1} \\ B_{h_2} \end{pmatrix} = \begin{pmatrix} D_{x_1} \\ D_{\bar{x}_2} \end{pmatrix} + D_{x_2} \cdot \begin{pmatrix} B_{\bar{x}_1} \\ B_{\bar{x}_2} \end{pmatrix}$
	$= \begin{pmatrix} D_{x_1} \\ D_{x_2} \end{pmatrix} + D_{\bar{x}_2} \cdot \begin{pmatrix} B_{\bar{x}_1} \\ B_{\bar{x}_2} \end{pmatrix}$
$\boldsymbol{B_p} = \boldsymbol{D_x} + D_{\bar{x}_2} \cdot \boldsymbol{B_x}$	$\boldsymbol{B_h} = \boldsymbol{D_x} + D_{\bar{x}_2} \cdot \boldsymbol{B_{\bar{x}}}$

Now, solving $\boldsymbol{B_h} = \boldsymbol{D_x} + D_{\bar{x}_2} \cdot \boldsymbol{B_{\bar{x}}}$ for $\boldsymbol{D_x}$ yields $\boldsymbol{D_x} = \boldsymbol{B_h} - D_{\bar{x}_2} \cdot \boldsymbol{B_{\bar{x}}}$. Finally, inserting $\boldsymbol{D_x}$ into $\boldsymbol{B_p}$ results in the functional relationship in demand

$$\begin{aligned} \boldsymbol{B_p} =& \boldsymbol{B_h} - D_{\bar{x}_2} \cdot \boldsymbol{B_{\bar{x}}} + D_{\bar{x}_2} \cdot \boldsymbol{B_x} \\ =& \boldsymbol{B_h} + D_{\bar{x}_2} (\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}}) \end{aligned}$$

and completes the proof. □

To conclude, Result 9 provides the solution of Problem 2 for simple models. In the following section, we extend our findings about the functional relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ to multiple models comprising $Q > 2$ auxiliary variables.

5.2.3.2 Multiple Models Comprising $Q > 2$ Auxiliary Variables

As a reminder, we define $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iQ})^T = (1, x_{i2}, \ldots, x_{iQ})^T$ as an auxiliary vector of the person-level model and $\boldsymbol{\bar{x}_i} = (\bar{x}_{i1} = 1, \bar{x}_{i2}, \ldots, \bar{x}_{iQ})^T$ as an auxiliary vector of the reduced person-level model. The step to relate the multiple coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ is the same as in Section 5.2.3.1. We start by deducing the form of the multiple coefficient vector $\boldsymbol{B_{\bar{x}}}$. The proof is significantly more elaborate as for the dimension $Q = 2$. Hence, we divide the proof into two parts. In the first part, we concentrate on the case of $Q = 3$ auxiliary variables. We show that already for $Q = 3$ auxiliaries, the formulas become cumbersome. Therefore, we propose an alternative proceeding to prove that the lemma is applicable for dimensions $Q > 2$ using partial regression arguments and the FWL theorem. Subsequently, we derive the multidimensional relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ by exploiting the form of $\boldsymbol{B_{\bar{x}}}$.

Lemma 4. ***The Form of the Coefficient $B_{\bar{x}}$ Resulting from Regressing the Original Auxiliaries on its Constructed Mean Values***
Let $q' = 2, \ldots, Q$ be an index without an intercept and let $q = 1, \ldots, Q$ be an index including an intercept. Suppose the vector $\boldsymbol{B}_{\bar{\boldsymbol{x}}'_{\boldsymbol{q}}} = (B_{\bar{x}_{1q'}}, B_{\bar{x}_{2q'}}, \ldots, B_{\bar{x}_{qq'}}, \ldots, B_{\bar{x}_{Qq'}})^T$ contains the coefficients arising from

$$\begin{aligned} x_{iq'} &= \boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{q'}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i}} + r_i^{B_{\bar{x}'_q}} \\ &= B_{\bar{x}_{1q'}} \bar{x}_{i1} + B_{\bar{x}_{2q'}} \bar{x}_{i2} + \ldots + B_{\bar{x}_{qq'}} \bar{x}_{iq} + \ldots + B_{\bar{x}_{QQ}} \bar{x}_{iQ} + r_i^{B_{\bar{x}_{q'}}} . \end{aligned} \tag{5.42}$$

Consider that all $(Q-1)$ coefficient vectors $\boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{q'}}}$ of dimension Q are summarized to a $Q \times (Q-1)$-matrix $\boldsymbol{B}_{\bar{\boldsymbol{x}}} = (\boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{1}}}, \ldots, \boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{q'}}}, \ldots, \boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{Q}}})^T$. Then $\boldsymbol{B}_{\bar{\boldsymbol{x}}}$ has the form

$$\boldsymbol{B}_{\bar{\boldsymbol{x}}} = \begin{pmatrix} B_{\bar{x}_{12}} & \ldots & B_{\bar{x}_{1q'}} & \ldots & B_{\bar{x}_{1Q}} \\ B_{\bar{x}_{22}} & \ldots & B_{\bar{x}_{2q'}} & \ldots & B_{\bar{x}_{2Q}} \\ \vdots & \ddots & & & \vdots \\ B_{\bar{x}_{q2}} & & B_{\bar{x}_{qq'}} & & B_{\bar{x}_{qQ}} \\ \vdots & & & \ddots & \vdots \\ B_{\bar{x}_{Q2}} & \ldots & B_{\bar{x}_{Qq'}} & \ldots & B_{\bar{x}_{QQ}} \end{pmatrix} = \begin{pmatrix} 0 & \ldots & \ldots & \ldots & 0 \\ 1 & 0 & \ldots & \ldots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & 1 \end{pmatrix} . \tag{5.43}$$

It should be remarked that the index q' of the variable of interest $x_{iq'}$ on the left-hand side of regression (5.42) runs from 2 to Q, excluding the intercept. The index of the explanatory variable $\bar{x}_{iq}$, in turn, on the right-hand side of regression (5.42) runs from 1 to Q, including the intercept. Therefore, $\boldsymbol{B}_{\bar{\boldsymbol{x}}}$ is of dimension $Q \times (Q-1)$.

Proof. **$\boldsymbol{B}_{\bar{\boldsymbol{x}}}$ in Case of $Q = 3$ Auxiliary Variables**

For $Q = 3$ auxiliaries, the coefficient matrix

$$\boldsymbol{B}_{\bar{\boldsymbol{x}}} = (\boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{2}}}, \boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{3}}})^T = \begin{pmatrix} B_{\bar{x}_{12}} & B_{\bar{x}_{13}} \\ B_{\bar{x}_{22}} & B_{\bar{x}_{23}} \\ B_{\bar{x}_{32}} & B_{\bar{x}_{33}} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{5.44}$$

is of dimension (3×2). The first column vector $\boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{2}}}$ results from the regression

$$\begin{aligned} x_{i2} &= \boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{2}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i}} + r_i^{B_{\bar{x}_2}} \\ &= B_{\bar{x}_{12}} x_{i1} + B_{\bar{x}_{22}} \bar{x}_{i2} + B_{\bar{x}_{32}} \bar{x}_{i3} + r_i^{B_{\bar{x}_2}}, \end{aligned}$$

where

$$\boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{2}}} = \left(\sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}{}^T \right)^{-1} \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{x}_{i2}.$$

The problem arising is the analytical row-by-row representation of the inverse of $\sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}{}^T$ in $\boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{2}}}$. The well-known Gauß-Jordan algorithm only provides a numerical solution. However, we are interested in an analytical solution. A possible remedy is constituted by the relation of the inverse of a matrix to its adjugate and its determinate. If A is a $(Q \times Q)$ invertible matrix, then the inverse of A is given by

$$A^{-1} = \frac{1}{\text{det}(A)} \text{adj}(A). \tag{5.45}$$

Thus, to determine the inverse of $\sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^T$, we have to compute its adjugate and determinate. At first, the adjugate of a (3×3) matrix

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

is described by

$$\begin{aligned}
\text{adj}(A) &= \begin{pmatrix} \text{det}\begin{pmatrix} e & f \\ h & i \end{pmatrix} & -\text{det}\begin{pmatrix} d & f \\ g & i \end{pmatrix} & \text{det}\begin{pmatrix} d & e \\ g & h \end{pmatrix} \\ -\text{det}\begin{pmatrix} b & c \\ h & i \end{pmatrix} & \text{det}\begin{pmatrix} a & c \\ g & i \end{pmatrix} & -\text{det}\begin{pmatrix} a & b \\ g & h \end{pmatrix} \\ \text{det}\begin{pmatrix} b & c \\ e & f \end{pmatrix} & -\text{det}\begin{pmatrix} a & c \\ d & f \end{pmatrix} & \text{det}\begin{pmatrix} a & b \\ d & e \end{pmatrix} \end{pmatrix}^T \\
&= \begin{pmatrix} ei - fh & fg - di & dh - eg \\ ch - bi & ai - cg & bg - ah \\ bf - ce & cd - af & ae - bd \end{pmatrix}^T \\
&= \begin{pmatrix} ei - fh & ch - bi & bf - ce \\ fg - di & ai - cg & cd - af \\ dh - eg & bg - ah & ae - bd \end{pmatrix}.
\end{aligned}$$

In our case, the adjugate simplifies to a triangular matrix, since $\sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^T$ is symmetric. It is given by

$$\text{adj}(\sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^T) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ & a_{22} & a_{23} \\ & & a_{33} \end{pmatrix}, \tag{5.46}$$

where

$$\begin{aligned}
a_{11} &= \sum_{i\in U_p} \bar{x}_{i2}^2 \sum_{i\in U_p} \bar{x}_{i3}^2 - \sum_{i\in U_p} \bar{x}_{i3}\bar{x}_{i2} \sum_{i\in U_p} \bar{x}_{i2}\bar{x}_{i3} \\
a_{12} &= \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i2}\bar{x}_{i3} - \sum_{i\in U_p} \bar{x}_{i2} \sum_{i\in U_p} \bar{x}_{i3}^2 \\
a_{13} &= \sum_{i\in U_p} \bar{x}_{i2} \sum_{i\in U_p} \bar{x}_{i3}\bar{x}_{i2} - \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i2}^2 \\
a_{22} &= \sum_{i\in U_p} 1^2 \sum_{i\in U_p} \bar{x}_{i3}^2 - \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i3} \\
a_{23} &= \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i2} - \sum_{i\in U_p} 1^2 \sum_{i\in U_p} \bar{x}_{i3}\bar{x}_{i2} \\
a_{33} &= \sum_{i\in U_p} 1^2 \sum_{i\in U_p} \bar{x}_{i2}^2 - \sum_{i\in U_p} \bar{x}_{i2} \sum_{i\in U_p} \bar{x}_{i2}.
\end{aligned}$$

Secondly, the determinate of a (3×3) matrix A can be obtained from Laplace expansion (or Sarrus's rule)

$$\text{det}(A) = aei + bfg + cdh - gec - hfa - idb.$$

Accordingly, the determinant of matrix $\sum_{i\in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^T$ is obtained by

$$\begin{aligned}
\text{det}(\sum_{i\in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^T) = &\sum_{i\in U_p} 1 \sum_{i\in U_p} \bar{x}_{i2}^2 \sum_{i\in U_p} \bar{x}_{i3}^2 + \sum_{i\in U_p} \bar{x}_{i2} \sum_{i\in U_p} \bar{x}_{i2}\bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i3} \\
&+ \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i2} \sum_{i\in U_p} \bar{x}_{i2}\bar{x}_{i3} - \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U} \bar{x}_{i2}^2 \sum_{i\in U_p} \bar{x}_{i3} \\
&- \sum_{i\in U_p} \bar{x}_{i2}\bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i2}\bar{x}_{i3} \sum_{i\in U_p} 1 - \sum_{i\in U_p} \bar{x}_{i3}^2 \sum_{i\in U_p} \bar{x}_{i2} \sum_{i\in U_p} \bar{x}_{i2}. \qquad (5.47)
\end{aligned}$$

Inserting the adjugate (5.46) and determinant (5.47) into formula (5.45), we obtain the inverse of the matrix $\sum_{i\in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i}} \bar{\boldsymbol{x}}_{\boldsymbol{i}}^T$.

Now, given the row-by-row representation of the inverse, we start to prove that the first column vector $\boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{2}}}$ in (5.44) equals

$$\begin{aligned}
\boldsymbol{B}_{\bar{\boldsymbol{x}}_{\boldsymbol{2}}} &= \begin{pmatrix} \sum\limits_{i\in U_p} 1^2 & \sum\limits_{i\in U_p} \bar{x}_{i2} & \sum\limits_{i\in U_p} \bar{x}_{i3} \\ \sum\limits_{i\in U_p} \bar{x}_{i2} & \sum\limits_{i\in U_p} \bar{x}_{i2}^2 & \sum\limits_{i\in U_p} \bar{x}_{i2}\bar{x}_{i3} \\ \sum\limits_{i\in U_p} \bar{x}_{i3} & \sum\limits_{i\in U_p} \bar{x}_{i3}\bar{x}_{i2} & \sum\limits_{i\in U_p} \bar{x}_{i3}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum\limits_{i\in U_p} x_{i2} \\ \sum\limits_{i\in U_p} \bar{x}_{i2}x_{i2} \\ \sum\limits_{i\in U_p} \bar{x}_{i3}x_{i2} \end{pmatrix} \\
&= \begin{pmatrix} B_{\bar{x}_{12}} \\ B_{\bar{x}_{22}} \\ B_{\bar{x}_{32}} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.
\end{aligned}$$

To show that $B_{\bar{x}_{22}} = 1$, we have to multiply the second line of adj$(\sum_{i\in U_p} \boldsymbol{\bar{x}_i \bar{x}_i})$ in (5.46) by $\sum_{i\in U_p} \boldsymbol{\bar{x}_i \bar{x}_{i2}}$ given by

$$\begin{aligned}
&(\sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i2}\bar{x}_{i3} - \sum_{i\in U_p} \bar{x}_{i2} \sum_{i\in U_p} \bar{x}_{i3}^2) \cdot \sum_{i\in U_p} 1 \cdot x_{i2} \\
&\qquad + (\sum_{i\in U_p} 1^2 \sum_{i\in U_p} \bar{x}_{i3}^2 - \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i3}) \cdot \sum_{i\in U_p} \bar{x}_{i2} x_{i2} \\
&\qquad + (\sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i2} - \sum_{i\in U_p} 1^2 \sum_{i\in U_p} \bar{x}_{i3}\bar{x}_{i2}) \cdot \sum_{i\in U_p} \bar{x}_{i3} x_{i2} \\
&\quad = \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i2}\bar{x}_{i3} \cdot \sum_{i\in U_p} x_{i2} - \sum_{i\in U_p} \bar{x}_{i2} \sum_{i\in U_p} \bar{x}_{i3}^2 \cdot \sum_{i\in U_p} x_{i2} \\
&\qquad + \sum_{i\in U_p} 1^2 \sum_{i\in U_p} \bar{x}_{i3}^2 \cdot \sum_{i\in U_p} \bar{x}_{i2} x_{i2} - \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i3} \cdot \sum_{i\in U_p} \bar{x}_{i2} x_{i2} \\
&\qquad + \sum_{i\in U_p} \bar{x}_{i3} \sum_{i\in U_p} \bar{x}_{i2} \cdot \sum_{i\in U_p} \bar{x}_{i3} x_{i2} - \sum_{i\in U_p} 1^2 \sum_{i\in U_p} \bar{x}_{i3}\bar{x}_{i2} \cdot \sum_{i\in U_p} \bar{x}_{i3} x_{i2}. \qquad (5.48)
\end{aligned}$$

Since $\sum_{i\in U_p} x_{i2} = \sum_{i\in U_p} \bar{x}_2$ and $\sum_{i\in U_p} \bar{x}_2 x_{i2} = \sum_{i\in U_p} \bar{x}_{i2}^2$ equation (5.48) is equal to the determinant in (5.47). Therefore, it follows that $B_{\bar{x}_{22}} = 1$.

In a similar manner, it can be shown that the first and the third row of adj$(\sum_{i\in U_p} \boldsymbol{\bar{x}_i \bar{x}_i}^T)$ in (5.46) multiplied by $\sum_{i\in U_p} \boldsymbol{\bar{x}_i \bar{x}_{i1}}$ equals zero. Therefore, it is valid that $B_{\bar{x}_{12}} = B_{\bar{x}_{32}} = 0$.

The analogous proceeding can by employed to prove that $\boldsymbol{B_{\bar{x}_3}} = (B_{13}, B_{23}, B_{33})^T = (0, 0, 1)^T$. Thereby, (5.43) is proven for the case of $Q = 3$ auxiliaries.

Unfortunately, in the case of $Q > 3$ auxiliary variables the representation of the inverse of $\sum_{i\in U_p} \boldsymbol{\bar{x}_i \bar{x}_i}^T$ becomes cumbersome. Following the Laplace expansion, the determinant of a $(Q \times Q)$ matrix A is expressed by a weighted sum of determinants of Q submatrices of A of size $(Q-1) \times (Q-1)$. This implies that already for $Q = 4$ auxiliary variables, we have to solve 4 submatrices of $\sum_{i\in U_p} \boldsymbol{\bar{x}_i \bar{x}_i}^T$ of size (3×3). The computational effort increases disproportionally with Q. To handle this obstacle, we aim at reducing the dimension of the matrix on which the inverse is applied.

$\boldsymbol{B_{\bar{x}}}$ in Case of $Q > 3$ Auxiliary Variables

Since the problem is the analytical representation of the inverse, the idea is to reduce the dimension of the matrix on which the inverse is applied without the reduction of Q itself. A solution provides the concept of partial regression and the FWL theorem, presented in Section 5.1.2.2.

For a better comprehension of the formulas, we once again express matrix $\boldsymbol{B_{\bar{x}}}$

$$\boldsymbol{B_{\bar{x}}} = \begin{pmatrix} B_{\bar{x}_{12}} & \dots & B_{\bar{x}_{1q'}} & \dots & B_{\bar{x}_{1Q}} \\ B_{\bar{x}_{22}} & \dots & B_{\bar{x}_{2q'}} & \dots & B_{\bar{x}_{2Q}} \\ \vdots & \ddots & & & \vdots \\ B_{\bar{x}_{q2}} & & B_{\bar{x}_{qq'}} & & B_{\bar{x}_{qQ}} \\ \vdots & & & \ddots & \vdots \\ B_{\bar{x}_{Q2}} & \dots & B_{\bar{x}_{Qq'}} & \dots & B_{\bar{x}_{QQ}} \end{pmatrix} = \begin{pmatrix} 0 & \dots & \dots & \dots & 0 \\ 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}.$$

All coefficients within $\boldsymbol{B}_{\bar{x}}$ results from Q regressions

$$\begin{aligned} x_{iq'} &= \boldsymbol{B}_{\bar{\boldsymbol{x}}'_{\boldsymbol{q}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i}} + r_i^{B_{\bar{x}'_q}} \quad \text{for } q' = 2, \dots, Q \\ &= B_{\bar{x}_{1q'}} \bar{x}_{i1} + B_{\bar{x}_{2q'}} \bar{x}_{i2} + \dots + B_{\bar{x}_{qq'}} \bar{x}_{iq} + \dots + B_{\bar{x}_{QQ}} \bar{x}_{iQ} + r_i^{B_{\bar{x}'_q}}. \end{aligned}$$

Now, we divide the coefficient matrix $\boldsymbol{B}_{\bar{x}}$ into three parts: the first row, the diagonal, and the minor diagonal. In order to prove Lemma 4, we have to show that

a) all first row elements $B_{\bar{x}_{1q'}} = 0$ for $q' = 2, \dots, Q$,

b) all diagonal elements $B_{\bar{x}_{qq'}} = 1$ for $q = 1, \dots, Q$, $q = q'$ and $q \neq 1$, and

c) all minor diagonal elements $B_{\bar{x}_{qq'}} = 0$ for $q \neq q'$ and $q \neq 1$.

a) We start by examining the coefficients on the diagonal

$$B_{\bar{x}_{qq'}} = \left(\sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{iq}} \bar{\boldsymbol{x}}_{\boldsymbol{iq}}{}^T \right)^{-1} \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{iq}} x_{iq'} = 1 \;\text{ for } q = q' \;\text{ and } q \neq 1. \tag{5.49}$$

That are all coefficients concerning the auxiliary and household mean value of the same variable. In order to reduce the dimension of matrix $\sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{iq}} \bar{\boldsymbol{x}}_{\boldsymbol{iq}}{}^T$ in (5.49), which has to be inverted, we apply the FWL theorem. As pointed out in Section 5.1.2.2, the FWL states that in a multiple regression the coefficient of any specific single variable can also be obtained by first partialing out the effects of all other explanatory variables from both the specific single variable and the variable of interest, and subsequently regressing the remaining variation of the variable of interest on the remaining variation of the explanatory variables. According to this, we have to partial out the effects of all Q household mean values except from the q-th variable from both the q-th original variable $x_{iq'}$ and its q-th mean value $\bar{x}_{iq}$. For this purpose, we define $\bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}$ as vector without the q-th element of dimension $(Q-1)$. Then, to partial out the effects of all other explanatory variables except for the q-th variable from $x_{iq'}$ and $\bar{x}_{iq}$, we have to determine the following two regressions

$$\left. \begin{aligned} x_{iq'} &= \boldsymbol{H}_{\bar{x}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} + r_i^{H_{\bar{x}}} \\ \bar{x}_{iq} &= \boldsymbol{K}_{\bar{x}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} + r_i^{K_{\bar{x}}} \end{aligned} \right. \quad \text{for } q = q' \;\text{ and } q \neq 1 \tag{5.50}$$

with $\boldsymbol{H}_{\bar{x}}$ and $\boldsymbol{K}_{\bar{x}}$ as vectors of dimension $(Q-1)$ in obvious notation. Both variables of interest on the left-hand side of the regressions (5.50) concern the same auxiliary. For explanation, consider $x_{iq'}$ is the variable `sex`. Then, $\bar{x}_{iq}$ is determined by the household mean value of `sex`. The explanatory variable $\bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}$ on the right-hand side of (5.50) contains the household mean values of all remaining variables except for that of `sex`.

Applying the FWL theorem, the coefficients on the diagonal $B_{\bar{x}_{qq'}}$ can be calculated by regressing the residuals $r_i^{H_{\bar{x}}}$ on the residuals $r_i^{K_{\bar{x}}}$ obtained from (5.50)

$$r_i^{H_{\bar{x}}} = B_{\bar{x}_{qq'}} r_i^{K_{\bar{x}}} + r_i^{B_{\bar{x}q}},$$

where

$$B_{\bar{x}_{qq'}} = \frac{\sum_{i \in U_p} r_i^{K_{\bar{x}}} r_i^{H_{\bar{x}}}}{\sum_{i \in U_p} (r_i^{K_{\bar{x}}})^2} \quad \text{for } q = q' \text{ and } q \neq 1. \tag{5.51}$$

As a result, we have successfully reduced the dimension of the matrix that have to be inverted from $(Q \times Q)$ of $(\sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{iq}} \bar{\boldsymbol{x}}_{\boldsymbol{iq}}{}^T)$ in (5.49) to the dimension (1×1) of $r_i^{K_{\bar{x}}}$ in (5.51).

It remains to show that the numerator and denominator of (5.51) are equal. It is valid that

$$\begin{aligned}
\sum_{i \in U_p} r_i^{H_{\bar{x}}} r_i^{K_{\bar{x}}} &= \sum_{i \in U_p} (x_{iq'} - \boldsymbol{H}_{\bar{\boldsymbol{x}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}})(\bar{x}_{iq} - \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}) \\
&= \sum_{i \in U_p} x_{iq'} \bar{x}_{iq} - \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^T \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} x_{iq'} - \boldsymbol{H}_{\bar{\boldsymbol{x}}}{}^T \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} \bar{x}_{iq} \\
&\quad + \boldsymbol{H}_{\bar{\boldsymbol{x}}}{}^T \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T \boldsymbol{K}_{\bar{\boldsymbol{x}}}.
\end{aligned} \tag{5.52}$$

Equation (5.52) can be simplified by the equivalence of

$$\begin{aligned}
\boldsymbol{H}_{\bar{\boldsymbol{x}}} &= \sum_{i \in U_p} (\bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T)^{-1} \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T x_{iq} \\
&= \sum_{i \in U_p} (\bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T)^{-1} \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T \bar{x}_{iq} \\
&= \boldsymbol{K}_{\bar{\boldsymbol{x}}}.
\end{aligned}$$

The equality of $\sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T x_{iq} = \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T \bar{x}_{iq}$ follows directly from the equality of the totals $\sum_{i \in U_p} \bar{x}_{iq} = \sum_{i \in U_p} x_{iq}$.

Hence, with $\boldsymbol{H}_{\bar{\boldsymbol{x}}} = \boldsymbol{K}_{\bar{\boldsymbol{x}}}$ Equation (5.52) becomes

$$\begin{aligned}
&= \sum_{i \in U_p} \bar{x}_{iq}^2 - 2\boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^T \sum_{i \in U_p} \bar{x}_{iq} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} + \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^T \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T \boldsymbol{K}_{\bar{\boldsymbol{x}}} \\
&= \sum_{g \in U_p} (\bar{x}_{iq} - \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}})^2 \\
&= \sum_{i \in U_p} r_i^{K_{\bar{x}}\,2}.
\end{aligned} \tag{5.53}$$

The equality of $\sum_{i \in U_p} \bar{x}_{iq}^2 = \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T x_{iq}$ is already shown in the proof of Lemma 3 in (5.39). Therefore, we have successfully proven the equality of the numerator and denominator in (5.51) and thus that all diagonal elements $B_{\bar{x}_{qq'}} = 1$ for all $q = q'$and $q \neq 1$.

b) We continue to show that all minor diagonal elements

$$B_{\bar{x}_{qq'}} = (\sum_{i \in U_p} \bar{x}_{iq} \bar{x}_{iq}^T)^{-1} \sum_{i \in U_p} \bar{x}_{iq} x_{iq'} = 0 \text{ for } q \neq q' \text{ and } q \neq 1.$$

For this purpose, we have to partial out the effects of all other explanatory variables from $x_{iq'}$ and $\bar{x}_{iq}$ by the following two regressions

$$\begin{aligned} x_{iq'} &= \tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}{}^{T} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}} + r_i^{\tilde{H}_{\bar{x}}} \\ \bar{x}_{iq} &= \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^{T} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}} + r_i^{K_{\bar{x}}} \end{aligned} \quad \text{for } q' \neq q \text{ and } q \neq 1. \tag{5.54}$$

The only difference to the aforementioned regressions in (5.50) is that the variable of interest in the upper regression refers to a different variable than the variable of interest in the lower regression. Thus, we denote the coefficient in the upper regression as $\tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}$ to differentiate it from $\boldsymbol{H}_{\bar{\boldsymbol{x}}}$ in (5.50). The explanatory variables and the complete lower regression are the same.

As before, we obtain the coefficient $B_{\bar{x}_{qq'}}$ by regressing $r_i^{\tilde{H}_{\bar{x}}}$ on $r_i^{K_{\bar{x}}}$

$$r_i^{\tilde{H}_{\bar{x}}} = B_{\bar{x}_{qq'}} r_i^{K_{\bar{x}}} + r_i^{B_{\bar{x}_q}},$$

where

$$B_{\bar{x}_{qq'}} = \frac{\sum_{i \in U_p} r_i^{K_{\bar{x}}} r_i^{\tilde{H}_{\bar{x}}}}{\sum_{i \in U_p} (r_i^{K_{\bar{x}}})^2} \quad \text{for } q' \neq q \text{ and } q \neq 1. \tag{5.55}$$

It remains to show that $B_{\bar{x}_{qq'}} = 0$ in (5.55). We know from (5.53) that the denominator $\sum_{i \in U_p} r_i^{K_{\bar{x}}2} \neq 0$. Consequently, we analyze the numerator $\sum_{i \in U_p} r_i^{\tilde{H}_{\bar{x}}} r_i^{K_{\bar{x}}}$ given by

$$\begin{aligned} \sum_{i \in U_p} r_i^{\tilde{H}_{\bar{x}}} r_i^{K_{\bar{x}}} &= \sum_{i \in U_p} (x_{iq'} - \tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}{}^{T} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}})(\bar{x}_{iq} - \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^{T} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}}) \\ &= \sum_{i \in U_p} x_{iq'} \bar{x}_{iq} - \sum_{i \in U_p} x_{iq'} \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^{T} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}} - \sum_{i \in U_p} \bar{x}_{iq} \tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}{}^{T} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}} \\ &\quad + \tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}{}^{T} \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}}{}^{T} \boldsymbol{K}_{\bar{\boldsymbol{x}}}. \end{aligned} \tag{5.56}$$

The second term in (5.56) can be rewritten as

$$\begin{aligned} \sum_{i \in U_p} x_{iq'} \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^{T} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}} &= \sum_{i \in U_p} x_{iq'} (\bar{x}_{iq} - r_i^{K_{\bar{x}}}) \\ &= \sum_{i \in U_p} x_{iq'} \bar{x}_{iq} - \underbrace{\sum_{i \in U_p} x_{iq'} r_i^{K_{\bar{x}}}}_{=0}. \end{aligned}$$

Note that $\sum_{i \in U_p} x_{iq'} r_i^{K_{\bar{x}}} = \sum_{i \in U_p} (\bar{x}_{iq} - \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^{T} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}}) x_{iq'} = 0$, since it constitutes the minimization problem of the lower regression in (5.54), which equals zero following the least squares theory (cf. Greene, 2003, Section 6.4; Wooldridge, 2013, Section 3.2).

In a similar manner, the third term in (5.56) can be rearranged to

$$\begin{aligned} \sum_{i \in U_p} \bar{x}_{iq} \tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}{}^{T} \bar{\boldsymbol{x}}_{\boldsymbol{i},-\boldsymbol{q}} &= \sum_{i \in U_p} \bar{x}_{iq} (x_{iq'} - r_i^{\tilde{H}_{\bar{x}}}) \\ &= \sum_{i \in U_p} \bar{x}_{iq} x_{iq'} - \underbrace{\sum_{i \in U_p} \bar{x}_{iq} r_i^{\tilde{H}_{\bar{x}}}}_{=0}. \end{aligned}$$

The latter term $\sum_{i \in U_p} \bar{x}_{iq} r_i^{\tilde{H}_{\bar{x}}} = \sum_{i \in U_p} (x_{iq'} - \tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}) \bar{x}_{iq}$ is equal to zero for the same least squares argumentation given above, as it determines the minimization problem of the upper regression in (5.54).

Finally, it can also be shown that the fourth term in (5.56) equals the remaining terms

$$\begin{aligned}
\tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}{}^T \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T \boldsymbol{K}_{\bar{\boldsymbol{x}}} &= \sum_{i \in U_p} (x_{iq'} - r_i^{\tilde{H}_{\bar{x}}})(\bar{x}_{iq} - r_i^{K_{\bar{x}}}) \\
&= \sum_{i \in U_p} x_{iq'} \bar{x}_{iq} - \underbrace{\sum_{i \in U_p} x_{iq'} r_i^{K_{\bar{x}}}}_{=0} - \underbrace{\sum_{i \in U_p} r_i^{\tilde{H}_{\bar{x}}} \bar{x}_{iq}}_{=0} + \underbrace{\sum_{i \in U_p} r_i^{\tilde{H}_{\bar{x}}} r_i^{K_{\bar{x}}}}_{=0} .
\end{aligned}$$

To summarize, inserting these rearrangements into (5.56), we obtain

$$\begin{aligned}
\sum_{i \in U_p} r_i^{\tilde{H}_{\bar{x}}} r_i^{K_{\bar{x}}} &= \sum_{i \in U_p} x_{iq'} \bar{x}_{iq} - \sum_{i \in U_p} x_{iq'} \boldsymbol{K}_{\bar{\boldsymbol{x}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} - \sum_{i \in U_p} \bar{x}_{iq} \tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} \\
&\quad + \tilde{\boldsymbol{H}}_{\bar{\boldsymbol{x}}}{}^T \sum_{i \in U_p} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}}{}^T \boldsymbol{K}_{\bar{\boldsymbol{x}}} \\
&= \sum_{i \in U_p} x_{iq'} \bar{x}_{iq} - \sum_{i \in U_p} x_{iq'} \bar{x}_{iq} - \sum_{i \in U_p} x_{iq'} \bar{x}_{iq} + \sum_{i \in U_p} x_{iq'} \bar{x}_{iq} \\
&= 0.
\end{aligned}$$

Therefore, we successfully proved that all minor elements $B_{qq'} = 0$ for all $q \neq q'$ and $q \neq 1$.

c) It remains to prove that all elements on the first row of (5.43) concerning the intercept terms equal to zero

$$B_{\bar{x}_{1q'}} = (\sum_{i \in U_p} \bar{x}_{iq} \bar{x}_{iq}^T)^{-1} \sum_{i \in U_p} \bar{x}_{iq} x_{i1} = 0 \text{ for } q' = 2, \ldots, Q.$$

For this purpose, we have to partial out the effects of all other explanatory variables from $x_{iq'}$ and $\bar{x}_{i1}$ by the following two regressions

$$\begin{aligned}
x_{iq'} &= \boldsymbol{H}_{\bar{\boldsymbol{x}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} + r_i^{H_{\bar{x}}} \\
\bar{x}_{i1} &= \tilde{\boldsymbol{K}}_{\bar{\boldsymbol{x}}}{}^T \bar{\boldsymbol{x}}_{\boldsymbol{i,-q}} + r_i^{\tilde{K}_{\bar{x}}} .
\end{aligned} \tag{5.57}$$

These regressions differ from the regressions in (5.50) only with respect to the variable of interest $\bar{x}_{i1}$. We denote the coefficient in the lower regression as $\tilde{\boldsymbol{K}}_{\bar{\boldsymbol{x}}}$ to differentiate it from $\boldsymbol{K}_{\bar{\boldsymbol{x}}}$ in (5.50) and (5.54). On both sides of the lower regression in (5.57) emerge a constant term, $x_{i1} = \bar{x}_{i1} = 1$. Hence, the variable of interest $\bar{x}_{i1}$ is exactly explained through x_{i1} and no remaining correlation between the variable of interest and the explanatory variables is captured by the residuals, which results in $r_i^{\tilde{K}_{\bar{x}}} = 0$. Consequently, from the regression

$$r_i^{H_{\bar{x}}} = B_{\bar{x}_{1q'}} r_i^{\tilde{K}_{\bar{x}}} + r_i^{B_{\bar{x}q}}$$

it immediately follows that $B_{\bar{x}_{1q'}} = 0$.

As a result, applying arguments of partial regressions and the FWL theorem, we successfully proved that all first row elements and minor diagonal elements equal zero and that all diagonal elements equal 1. Therefore, Lemma 4 is proven. □

We learn from Lemma 3 that the only nonzero entities in $\boldsymbol{B_{\bar{x}}}$ are the coefficients $B_{\bar{x}_{qq'}}$ with $q = q'$ and $q \neq 0$. These are the diagonal elements determining the influence of $\bar{x}_{iq}$ on $x_{iq'}$ when the variable of interest and explanatory variable concern the same auxiliary variable holding all other variables constant. As argued for simple models, the reason for the form of $\boldsymbol{B_{\bar{x}}}$ is that the variation of the original auxiliary variables is exactly explained by the household mean values. It is important to note that Lemma 3 holds only when regressing $\bar{x}_{iq}$ on $x_{iq'}$. It does not hold when regressing $x_{iq'}$ on $\bar{x}_{iq}$.

We continue with deriving the relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ for multiple models. The proceeding is analogous to the simple model case. Thus, we exploit the fact that an overlap model comprises both auxiliaries $\boldsymbol{x_i}$ and $\boldsymbol{\bar{x}_i}$. Then, we decompose $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ into the same coefficients obtained from the overlap model. For the decomposition, we apply two mediation models and interpret the coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ as direct effects. As done before, for a better comprehension we apply tables within the proof. The proof is kept short, as it is analogous to the proof of Result 9.

Result 10. ***The Functional Relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ in Multiple Models***
The coefficient $\boldsymbol{B_p}$, resulting from the person-level model $y_i = \boldsymbol{B_p}^T\boldsymbol{x_i} + r_i^{B_p}$, can be expressed as

$$\boldsymbol{B_p} = \boldsymbol{B_h} + \boldsymbol{D_{\bar{x}}}(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}}), \tag{5.58}$$

where $\boldsymbol{B_h}$ results from the reduced person-level model $y_i = \boldsymbol{B_h}^T\boldsymbol{\bar{x}_i} + r_i^{B_h}$. Coefficient $E_{\bar{x}_1}$ arises from the overlap model $y_i = D_{x_1}x_{i1} + \boldsymbol{D_x}^T\boldsymbol{x_i} + \boldsymbol{D_{\bar{x}}}^T\boldsymbol{\bar{x}_i} + r_i^D$ describing the overlap between the person-level auxiliary information $\boldsymbol{x_i}$ and the household-level auxiliaries $\boldsymbol{\bar{x}_i}$. Coefficient vectors $\boldsymbol{B_x}$ and $\boldsymbol{B_{\bar{x}}}$ are obtained from the auxiliary models $\bar{x}_i = \boldsymbol{B_x}^T\boldsymbol{x_i} + r_i^{B_{\bar{x}}}$ and $x_i = \boldsymbol{B_{\bar{x}}}^T\boldsymbol{\bar{x}_i} + r_i^{B_x}$, respectively.

Proof. The overlap between the multiple auxiliary variables of a person-level and a reduced person-level model, $\boldsymbol{x_i}$ and $\boldsymbol{\bar{x}_i}$, respectively, is constituted by the following overlap model

$$y_i = D_{x_1}x_1 + \boldsymbol{D_x}^T\boldsymbol{x_i} + \boldsymbol{D_{\bar{x}}}^T\boldsymbol{\bar{x}_i} + r_i^D. \tag{5.59}$$

with $\boldsymbol{D_x} = (D_{x2}, \dots, D_{x_{q'}}, \dots, D_{xQ})^T$ and $\boldsymbol{D_{\bar{x}}} = (D_{\bar{x}2}, \dots, D_{\bar{x}_{q'}}, \dots, D_{\bar{x}Q})^T$ for $q' = 2, \dots, Q$ as coefficients.

To decompose $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$, we apply two mediation models, as introduced in Section 5.2.1.1, and interpret both coefficients at hand as indirect effects. The results of the decomposition and multiple auxiliary model are summarized in Table 5.5.

Table 5.5: Multiple Models under consideration to derive a relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$

Person-level model	**Reduced person-level model**
$y_i = \boldsymbol{B_p^T x_i} + r_i^{B_p}$ with $\boldsymbol{B_p} = (B_{p_1}, B_{p_2}, \ldots, B_{p_{q'}}, \ldots, B_{p_Q})^T$	$y_i = \boldsymbol{B_h^T \bar{x}_g} + r_i^{B_h}$ with $\boldsymbol{B_h} = (B_{h_1}, B_{h_2}, \ldots, B_{h_{q'}}, \ldots, B_{h_Q})^T$
The auxiliary regressions are given by	
$\bar{x}_{i2} = \boldsymbol{B_{x_2}^T x_i} + r_i^{B_{\bar{x}_2}}$ $\vdots$ $\bar{x}_{iq'} = \boldsymbol{B_{x_{q'}}^T x_i} + r_i^{B_{\bar{x}_{q'}}}$ $\vdots$ $\bar{x}_{iQ} = \boldsymbol{B_{x_Q}^T x_i} + r_i^{B_{\bar{x}_Q}}$ with $\boldsymbol{B_{x_{q'}}} = (B_{x_{1q'}}, B_{x_{2q'}}, \ldots, B_{x_{qq'}}, \ldots, B_{x_{Qq'}})^T$	$x_{i2} = \boldsymbol{B_{\bar{x}_2}^T \bar{x}_i} + r_i^{B_{x_2}}$ $\vdots$ $x_{iq'} = \boldsymbol{B_{\bar{x}^T_{q'}} \bar{x}_i} + r_i^{B_{x_{q'}}}$ $\vdots$ $x_{iQ} = \boldsymbol{B_{\bar{x}_Q}^T \bar{x}_i} + r_i^{B_{x_Q}}$ with $\boldsymbol{B_{\bar{x}_{q'}}} = (B_{\bar{x}_{1q'}}, B_{\bar{x}_{2q'}}, \ldots, B_{\bar{x}_{qq'}}, \ldots, B_{\bar{x}_{Qq'}})^T$
The original coefficients are decomposed into	
$\begin{pmatrix} B_{p_1} \\ B_{p_2} \\ \vdots \\ B_{p_q} \\ \vdots \\ B_{p_Q} \end{pmatrix} = \begin{pmatrix} D_{x_1} \\ D_{x_2} \\ \vdots \\ D_{x_q} \\ \vdots \\ D_{x_Q} \end{pmatrix} + D_{\bar{x}_2} \cdot \begin{pmatrix} B_{x_{21}} \\ B_{x_{22}} \\ \vdots \\ B_{x_{2q}} \\ \vdots \\ B_{x_{2Q}} \end{pmatrix} + \ldots + D_{\bar{x}_{q'}} \cdot \begin{pmatrix} B_{x_{Q1}} \\ B_{x_{Q2}} \\ \vdots \\ B_{x_{Qq}} \\ \vdots \\ B_{x_{QQ}} \end{pmatrix}$	$\begin{pmatrix} B_{h_1} \\ B_{h_2} \\ \vdots \\ B_{h_q} \\ \vdots \\ B_{h_Q} \end{pmatrix} = \begin{pmatrix} D_{\bar{x}_1} \\ D_{\bar{x}_2} \\ \vdots \\ D_{\bar{x}_q} \\ \vdots \\ D_{\bar{x}_Q} \end{pmatrix} + D_{x_2} \cdot \begin{pmatrix} B_{\bar{x}_{21}} \\ B_{\bar{x}_{22}} \\ \vdots \\ B_{\bar{x}_{2q}} \\ \vdots \\ B_{\bar{x}_{2Q}} \end{pmatrix} + \ldots + D_{x_{q'}} \cdot \begin{pmatrix} B_{\bar{x}_{Q1}} \\ B_{\bar{x}_{Q2}} \\ \vdots \\ B_{\bar{x}_{Qq}} \\ \vdots \\ B_{\bar{x}_{QQ}} \end{pmatrix}$

According to Table 5.5, the decomposed coefficients can be expressed by

$$\boldsymbol{B_p} = \begin{pmatrix} D_{x_1} \\ \boldsymbol{D_x} \end{pmatrix} + \boldsymbol{D_{\bar{x}}} \cdot \boldsymbol{B_x}$$

and

$$\boldsymbol{B_h} = \begin{pmatrix} D_{x_1} \\ \boldsymbol{D_{\bar{x}}} \end{pmatrix} + \boldsymbol{D_x} \cdot \boldsymbol{B_{\bar{x}}}.$$

Fortunately, the different positions of D_{x_1}, $\boldsymbol{D_x}$ and $\boldsymbol{D_{\bar{x}}}$ prevent us from solving $\boldsymbol{B_p}$ for $\boldsymbol{B_h}$. However, we can resort the array of the household-level coefficient $\boldsymbol{B_h}$ by exploiting the form of $\boldsymbol{B_{\bar{x}}}$ derived in Lemma 4. Resorting $\boldsymbol{B_h}$ results in

$$\begin{pmatrix} B_{h_1} \\ B_{h_2} \\ \vdots \\ B_{h_q} \\ \vdots \\ B_{h_Q} \end{pmatrix} = \begin{pmatrix} D_{x_1} \\ D_{\bar{x}_2} \\ \vdots \\ D_{\bar{x}_{q'}} \\ \vdots \\ D_{\bar{x}_Q} \end{pmatrix} + \begin{pmatrix} B_{\bar{x}_{12}} & \dots & B_{\bar{x}_{1q'}} & \dots & B_{\bar{x}_{1Q}} \\ B_{\bar{x}_{22}} & \dots & B_{\bar{x}_{2q'}} & \dots & B_{\bar{x}_{2Q}} \\ \vdots & \ddots & & & \vdots \\ B_{\bar{x}_{q2}} & & B_{\bar{x}_{qq'}} & & B_{\bar{x}_{qQ}} \\ \vdots & & & \ddots & \vdots \\ B_{\bar{x}_{Q2}} & \dots & B_{\bar{x}_{Qq'}} & \dots & B_{\bar{x}_{QQ}} \end{pmatrix} \cdot \begin{pmatrix} D_{x_2} \\ \vdots \\ D_{x_{q'}} \\ \vdots \\ D_{x_Q} \end{pmatrix}$$

$$= \begin{pmatrix} D_{x_1} \\ D_{\bar{x}_2} \\ \vdots \\ D_{\bar{x}_{q'}} \\ \vdots \\ D_{\bar{x}_Q} \end{pmatrix} + \begin{pmatrix} 0 & \dots & \dots & \dots & 0 \\ 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} D_{x_2} \\ \vdots \\ D_{x_{q'}} \\ \vdots \\ D_{x_Q} \end{pmatrix},$$

swapping the positions of $\boldsymbol{D_x}$ and $\boldsymbol{D_{\bar{x}}}$ yields

$$= \begin{pmatrix} D_{x_1} \\ D_{x_2} \\ \vdots \\ D_{x_{q'}} \\ \vdots \\ D_{x_Q} \end{pmatrix} + \begin{pmatrix} 0 & \dots & \dots & \dots & 0 \\ 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} D_{\bar{x}_2} \\ \vdots \\ D_{\bar{x}_{q'}} \\ \vdots \\ D_{\bar{x}_Q} \end{pmatrix}.$$

Hence, both coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ are functions of the same $\boldsymbol{\tilde{D}_x} = (D_{x_1}, \boldsymbol{D_x}^T)^T = (D_{x_1}, D_{x_2}, \dots, D_{x_{q'}}, \dots, D_{x_Q})^T$ and $\boldsymbol{D_{\bar{x}}} = (D_{\bar{x}_2}, \dots, D_{\bar{x}_{q'}}, \dots, D_{\bar{x}_Q})^T$. Then, it is valid that

$$\boldsymbol{B_p} = \boldsymbol{\tilde{D}_x} + \boldsymbol{B_x} \cdot \boldsymbol{D_{\bar{x}}}$$

and

$$\boldsymbol{B_h} = \boldsymbol{\tilde{D}_x} + \boldsymbol{B_{\bar{x}}} \cdot \boldsymbol{D_{\bar{x}}}.$$

Solving $\boldsymbol{B_h} = \tilde{\boldsymbol{D}}_{\boldsymbol{x}} + \boldsymbol{B}_{\bar{\boldsymbol{x}}} \cdot \boldsymbol{D}_{\bar{\boldsymbol{x}}}$ for $\tilde{\boldsymbol{D}}_{\boldsymbol{x}}$ yields $\tilde{\boldsymbol{D}}_{\boldsymbol{x}} = \boldsymbol{B_h} - \boldsymbol{B}_{\bar{\boldsymbol{x}}} \cdot \boldsymbol{D}_{\bar{\boldsymbol{x}}}$. Inserting $\tilde{\boldsymbol{D}}_{\boldsymbol{x}}$ into $\boldsymbol{B_p}$ yields the functional relationship in demand

$$\begin{aligned}\boldsymbol{B_p} =& \boldsymbol{B_h} - \boldsymbol{B}_{\bar{\boldsymbol{x}}} \cdot \boldsymbol{D}_{\bar{\boldsymbol{x}}} + \boldsymbol{B}_{\boldsymbol{x}} \cdot \boldsymbol{D}_{\bar{\boldsymbol{x}}} \\ =& \boldsymbol{B_h} + (\boldsymbol{B}_{\boldsymbol{x}} - \boldsymbol{B}_{\bar{\boldsymbol{x}}})\boldsymbol{D}_{\bar{\boldsymbol{x}}} \end{aligned} \tag{5.60}$$

and completes the proof. □

To conclude, Result 10 provides the solution for Problem 2 and describes the functional relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ for multiple models. Given Result 9, we are able to explain the difference between the variances of a person- and a reduced person-level GREG estimator in objective function (5.31).

5.2.4 Inserting the Relationship between $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$ into the Efficiency Comparison

In this section, we reproduce our results derived so far and infer the implications for the efficiency comparison of a person-level and an integrated household-level GREG estimator. As a short reminder, the objective function (5.31) determining the efficiency comparison is given by

$$\begin{aligned}&\left(\mathrm{V}(\hat{T}_y^{\mathrm{GREG}}) - \mathrm{V}(\hat{T}_y^{\mathrm{INT}})\right) \Big/ \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1} \\ &= \sum_{g \in U_h} (r_g^{B_p})^2 - \sum_{g \in U_h} (r_g^{B_h^\circ})^2 \\ &= \sum_{g \in U_h} (r_g^{B_p})^2 - \sum_{g \in U_h} (r_g^{B_h})^2 - \sum_{g \in U_h} (\tilde{r}_g^{B_{x0}^\circ})^2 + \sum_{g \in U_h} (\tilde{r}_g^{B_{x0}^\circ \cdot F_x})^2 \\ &= \underbrace{\sum_{g \in U_h} (y_g - \boldsymbol{B_p}^T \boldsymbol{x_g})^2 - \sum_{g \in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g})^2}_{\text{variance component I}} \\ &\qquad - \underbrace{\left(\sum_{g \in U_h} (y_g - B_{x0}^\circ x_{g0})^2 - \sum_{g \in U_h} (y_g - B_{x0}^\circ \cdot \boldsymbol{F_x}^T \boldsymbol{x_g})^2 \right)}_{\text{variance component II}} . \end{aligned} \tag{5.61}$$

Result 8 derived in Section 5.2.1 provides the solution of the problem of different dimensions of the auxiliary variables of a person-level and an integrated GREG estimator, and thus of the corresponding residuals $r_g^{B_p}$ and $r_g^{B_h^\circ}$ (line 2). Following, the variance of an integrated GREG estimator, $\sum_{g \in U_h} (r_g^{B_h^\circ})^2$, can be decomposed into the variance of a reduced household-level model, $\sum_{g \in U_h} (r_g^{B_h})^2$, which is of the same dimension as $\sum_{g \in U_h} (r_g^{B_p})^2$, and two remaining variances (line 3). The remaining variances (line 5) capture the effects of the intercept disregarded by the reduced household model. To quantify the disregarded effect, we introduce the pseudo-residuals $\tilde{r}_g^{B_{x0}^\circ}$ and $\tilde{r}_g^{B_{x0}^\circ \cdot F_x}$. Therefore, we can rearrange the objective function (5.61) into two variance components: the difference of variances of the same dimension, termed as **variance component I**, and the effects caused by the intercept, termed as **variance component II**. The sum of both variance components is denoted by **total difference**.

Now, we examine variance component I in more detail. To asses variance component I, Result 10 (or Result 9 for simple models) supplied a functional relationship between the coefficients $\boldsymbol{B_p}$ and $\boldsymbol{B_h}$. Inserting Result 10 into variance component I in (5.61), we obtain

$$
\begin{aligned}
&\sum_{g\in U_h} (y_g - \boldsymbol{B_p}^T \boldsymbol{x_g})^2 - \sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g})^2 \\
&\quad = \sum_{g\in U_h} (y_g - (\boldsymbol{B_h} + \boldsymbol{D_{\bar{x}}}(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}}))^T \boldsymbol{x_g})^2 - \sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g})^2 \\
&\quad = \sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g} - \boldsymbol{D_{\bar{x}}}(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}})^T \boldsymbol{x_g})^2 - \sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g})^2 \\
&\quad = \sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g})^2 - \sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g})^2 + \sum_{g\in U_h} (\boldsymbol{D_{\bar{x}}}(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}})^T \boldsymbol{x_g})^2 \\
&\qquad\qquad - 2\sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g}) \boldsymbol{D_{\bar{x}}}(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}})^T \boldsymbol{x_g} \\
&\quad = \sum_{g\in U_h} (\boldsymbol{D_{\bar{x}}}(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}})^T \boldsymbol{x_g})^2 - 2\underbrace{\sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g}) \boldsymbol{D_{\bar{x}}}(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}})^T \boldsymbol{x_g}}_{(III)} . \qquad (5.62)
\end{aligned}
$$

If homoscedasticity is assumed in the integrated model, and thus also in the reduced household-level model, which implies that $v_g = 1$, term (III) equals zero, since $\sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T \boldsymbol{x_g})^T \boldsymbol{x_g} = 0$ determines the first derivative of the minimization problem of a reduced household-level model (5.16). This is not the case if heteroscedasticity is assumed and thus $v_g = N_g^{-1}$. The former case describes the integrated model proposed by Nieuwenbroek (1993) and the latter case determines the integrated model introduced by Lemaître and Dufour (1987). See Sections 3.1.2 and 3.1.3 for details on the different integrated models.

Finally, inserting Result (5.62) into variance component (I) (5.61) and differentiating two cases, yields

$$
\begin{aligned}
&\Big(\mathrm{V}(\hat{T}_y^{\mathrm{GREG}}\Big) - \mathrm{V}(\hat{T}_y^{\mathrm{INT}})\Big) \Big/ \frac{M^2}{m}\Big(1 - \frac{m}{M}\Big)(M-1)^{-1} \\
&\qquad = \sum_{g\in U_h} (r_g^{B_p})^2 - \sum_{g\in U_h} (r_g^{B_h^\circ})^2
\end{aligned}
$$

case a) if homoscedasticity is assumed,

$$
\begin{aligned}
&= \underbrace{\sum_{g\in U_h} (\boldsymbol{D_{\bar{x}}}(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}})^T \boldsymbol{x_g})^2}_{\text{I - Reduced difference of same dimension}} \\
&\qquad - \underbrace{\Big(\sum_{g\in U_h} (y_g - B^\circ_{x0} x_{g0})^2 - \sum_{g\in U_h} (y_g - B^\circ_{x0} \cdot \boldsymbol{F_x}^T \boldsymbol{x_g})^2\Big)}_{\text{II - Effects of the intercept}}, \qquad (5.63)
\end{aligned}
$$

case b) if heteroscedasticity is assumed,

$$= \underbrace{\sum_{g\in U_h} \left(\boldsymbol{D}_{\bar{\boldsymbol{x}}}(\boldsymbol{B}_{\boldsymbol{x}} - \boldsymbol{B}_{\bar{\boldsymbol{x}}})^T \boldsymbol{x}_{\boldsymbol{g}}\right)^2 - 2\sum_{g\in U_h}(y_g - \boldsymbol{B}_{\boldsymbol{h}}{}^T\boldsymbol{x}_{\boldsymbol{g}})\boldsymbol{D}_{\bar{\boldsymbol{x}}}(\boldsymbol{B}_{\boldsymbol{x}} - \boldsymbol{B}_{\bar{\boldsymbol{x}}})^T}_{\text{I - Reduced difference of same dimension}}$$
$$- \underbrace{\left(\sum_{g\in U_h}(y_g - B^{\circ}_{x0}x_{g0})^2 - \sum_{g\in U_h}(y_g - B^{\circ}_{x0}\cdot \boldsymbol{F}_{\boldsymbol{x}}{}^T\boldsymbol{x}_{\boldsymbol{g}})^2\right)}_{\text{II - Effects of the intercept}}. \tag{5.64}$$

Therefore, the total difference between the variance of a person-level and an integrated household-level GREG estimator is determined by two variance components describing

I) the difference of the variances of a person-level and a reduced household-level GREG estimator and

II) the effect of the intercept on the variance of an integrated GREG estimator ignored by the reduced household-level model.

Variance component II was completely neglected in the theorem (5.1) given by Steel and Clark (2007). In case a), variance component I is always positive, which implies that the variance of a person-level GREG estimator exceeds the one of a reduced household-level GREG estimator. This result is in accordance with the finding given in Section 5.1.2.1 that under single-stage cluster sampling, the variance is optimized by an unweighted coefficient depending on the aggregates of both the auxiliaries and variable of interest. In contrast, in case b) variance component I can be either positive or negative. The following two sections analyze variance components I and II in more detail to deduce predictions about their impact on the total difference in objective functions (5.63) and (5.64).

5.2.4.1 Variance Component I - Reduced Difference of the Same Dimension

Variance component I in objective function (5.63) or (5.64) is driven mainly by

i) $\boldsymbol{D}_{\bar{\boldsymbol{x}}}$

ii) $(\boldsymbol{B}_{\boldsymbol{x}} - \boldsymbol{B}_{\bar{\boldsymbol{x}}})$

iii) $\boldsymbol{B}_{\boldsymbol{h}}$ (only in case b).

We start by analyzing term i). $\boldsymbol{D}_{\bar{\boldsymbol{x}}}$ arises from the overlap model (5.59) and describes the effect of the constructed household mean values $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$ on the variable of interest y_i controlled for the original auxiliaries $\boldsymbol{x}_{\boldsymbol{i}}$. In other words, the higher the correlation between $\boldsymbol{x}_{\boldsymbol{i}}$ and $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$, the lower the variation of y_i not explained by $\boldsymbol{x}_{\boldsymbol{i}}$, and consequently the lower $\boldsymbol{D}_{\bar{\boldsymbol{x}}}$. To visualize the effect of $\boldsymbol{D}_{\bar{\boldsymbol{x}}}$ on the variance component I, we once more utilize Venn diagrams extensively discussed in Section 5.1.2.2. Typically, Venn diagrams address the case of having one or two explanatory variables. In the case of three explanatory variables, a simplex representation would be needed to draw Venn diagrams. Therefore, consider the following one-dimensional overlap model

$$y_i = D_x x_i + D_{\bar{x}} \bar{x}_i + r_i^D .$$

The left Venn diagram in Figure 5.12 illustrates the case of a high correlation between the auxiliaries, indicated by a large intersection between the circles of $\bar{x}_i$ and x_i. The right Venn diagram, in turn, depicts the case of a low correlation between $\bar{x}_i$ and x_i, resulting in a small intersection between the circles. Comparing these Venn diagrams, it becomes apparent that the higher the correlation between $\bar{x}_i$ and x_i, the lower $D_{\bar{x}}$. The reason is the variation common to both auxiliaries decreases. Of course, the result can easily be extended to the multiple variables. To summarize, we presume that variance component I is significantly affected by the correlation between the original auxiliaries, $\boldsymbol{x_i}$, and the constructed household mean values, $\boldsymbol{\bar{x}_i}$.

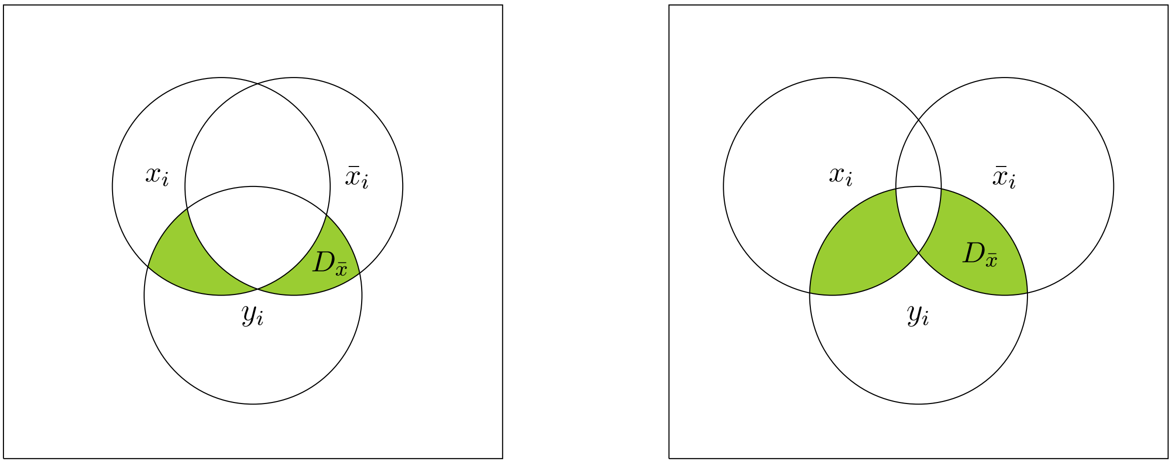

Figure 5.12: Venn diagram illustrating $D_{\bar{x}}$ with high (left side) and low (right side) correlation between x_i and $\bar{x}_i$

We continue with exploring term ii). It is noteworthy that $\boldsymbol{B_x}$ and $\boldsymbol{B_{\bar{x}}}$ are independent from the variable of interest y_i. The coefficients are influenced only by the auxiliary variables $\boldsymbol{x_i}$ and $\boldsymbol{\bar{x}_i}$. The coefficient $\boldsymbol{B_x}$, obtained from the auxiliary model

$$\boldsymbol{\bar{x}_i} = \boldsymbol{B_x}^T \boldsymbol{x_i} + r_i^{B_x},$$

describes the effect of the constructed household mean values on the original auxiliaries. The higher the correlation between $\boldsymbol{x_i}$ and $\boldsymbol{\bar{x}_i}$, the higher $\boldsymbol{B_x}$. Hence, $\boldsymbol{B_x}$ behaves contrarily to $\boldsymbol{D_{\bar{x}}}$, discussed in the previous paragraph. $\boldsymbol{B_{\bar{x}}}$ arises from the auxiliary model

$$\boldsymbol{x_i} = \boldsymbol{B_{\bar{x}}}^T \boldsymbol{\bar{x}_i} + r_i^{B_{\bar{x}}}.$$

Nevertheless, $\boldsymbol{B_{\bar{x}}}$ is independent from the correlation between $\boldsymbol{x_i}$ and $\boldsymbol{\bar{x}_i}$. Exploiting its special form, described by Lemma 4, we get

$$\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}} = \begin{pmatrix} B_{x_{12}} & \cdots & B_{x_{1q'}} & \cdots & B_{x_{1Q}} \\ B_{x_{22}} & \cdots & B_{x_{2q'}} & \cdots & B_{x_{2Q}} \\ \vdots & \ddots & & & \vdots \\ B_{x_{q2}} & & B_{x_{qq'}} & & B_{x_{qQ}} \\ \vdots & & & \ddots & \vdots \\ B_{x_{Q2}} & \cdots & B_{x_{Qq'}} & \cdots & B_{x_{QQ}} \end{pmatrix} - \begin{pmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} B_{x_{12}} & \cdots & B_{x_{1q'}} & \cdots & B_{x_{1Q}} \\ B_{x_{22}} - 1 & \cdots & B_{x_{2q'}} & \cdots & B_{x_{2Q}} \\ \vdots & \ddots & & & \vdots \\ B_{x_{q2}} & & B_{x_{qq'}} - 1 & & B_{x_{qQ}} \\ \vdots & & & \ddots & \vdots \\ B_{x_{Q2}} & \cdots & B_{x_{Qq'}} & \cdots & B_{x_{QQ}} - 1 \end{pmatrix}.$$

Hence, term ii) $(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}})$ is primarily constituted by $\boldsymbol{B_x}$. Only from the diagonal elements, comprising the coefficients concerning the same variable, an one is subtracted.

In case b), variance component I in (5.64) further depends on term iii). The extent of the household-level coefficient $\boldsymbol{B_h}$ is influenced by the correlation between the variable of interest y_g and the auxiliary variable $\boldsymbol{x_g}$.

To conclude, we expect that $(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}})$ and $\boldsymbol{D_{\bar{x}}}$ behave in contrary ways with respect to the correlation between $\boldsymbol{x_i}$ and $\boldsymbol{\bar{x}_i}$. Therefore, we diverge from Steel and Clark (2007, p. 54), who claimed that the difference of the variance of a person-level and an integrated GREG estimator (given in their second theorem see (5.1)) "[...] depends on the extent to which $\boldsymbol{\bar{x}}_i$ helps to predict y_i after $\boldsymbol{x_i}$ has already been controlled for, i.e., the extent to which a linear contextual effect helps to predict $r_i^{B_c}$ over $i \in U_p$, using a weighted least squares regression weighted by N_g." This statement is equivalent with the following: the higher the correlation between $\boldsymbol{x_i}$ and $\boldsymbol{\bar{x}_i}$, the lower the difference between the variances of a person-level and an integrated GREG estimator. Therefore, the statement considers only the effect of i) $\boldsymbol{D_{\bar{x}}}$, but not of ii) $(\boldsymbol{B_x} - \boldsymbol{B_{\bar{x}}})$.

5.2.4.2 Variance Component II - Effects of the Intercept

Variance component II is equal in objective functions (5.63) and (5.64). It captures the effect of separating the intercept from the integrated model and depends on

i) $B_{x_0}^{\circ}$ and

ii) $\boldsymbol{F_x}$.

We start by explaining term i). $B^{\circ}_{x_0}$ arises from the integrated household-level model

$$y_g = B^{\circ}_{x_0} x_{g0} + \boldsymbol{B}^{\circ T}_{\boldsymbol{x}} \boldsymbol{x_g} + r^{B^{\circ}}_g$$

and accounts for the effect of the intercept on y_g when controlling for the remaining auxiliaries $\boldsymbol{x_g}$. The interpretation of the intercept in a regression model should be treated with caution. Geometrically speaking, the intercept indicates the intersection of the regression hyperplane with the ordinate. As such, the intercept gives the value of the variables of interest when all other explanatory variables are set to zero. The interpretation is not admissible if the zero lies outside the range of the observed data. Thus, the intercept should be interpreted only from a technical point of view (cf. von Auer, 2007, p. 61). The intercept affects a shift of the regression hyperplane such that the residuals do not have an overall positive or negative bias. Hence, it is difficult to predict on which factors the intercept $B^{\circ}_{x_0}$ depends.

We continue with discussing term ii). $\boldsymbol{F_x}$ is obtained from the auxiliary model

$$x_{0g} = \boldsymbol{F_x}^T \boldsymbol{x_g} + r^{F_x}_g$$

and describes the influence of the auxiliaries $\boldsymbol{x_g}$ on a constant, x_{0g}. In Section 5.2.1.1, we already derived that $\boldsymbol{F_x} = (\sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{x_g}^T)^{-1} \boldsymbol{x_g}$ is independent from the variable of interest. Instead $\boldsymbol{F_x}$ is affected only by the explanatory variables.

To conclude, a prediction about the impact of variance component II on the total difference is difficult, because the intercept $B^{\circ}_{x_0}$ is of a more technical nature, and $\boldsymbol{F_x}$ depends only on the auxiliaries $\boldsymbol{x_g}$.

5.2.5 Simulation Study

In order to explore the previously discussed presumptions on variance components I and II, we run a MC simulation study. The simulation study is based on the same simulation setup as introduced in Section 3.4.1. The presented results focus on case b) with $v_i = 1$ and objective function (5.64). The results for case a) are very similar and can be found in Appendix C. One thousand samples of $m = 1500$ households are drawn via simple random sampling. The auxiliary variables are presented in Table 3.1. As variable of interest, we choose `inc`, as defined in Table 3.7.

We start by contrasting the total difference in (5.64) against variance components I and II. Figure 5.13 makes apparent that variance component I increases with the total difference, whereas variance component II decreases. The red lines divided the plots into quadrants. Within the lower left quadrant, all samples have negative signs for both variance components. Within the lower right quadrant, all samples points show a negative sign for variance component II and a positive sign for variance component I. It can be seen that the sign of the total difference and the first variance component I coincide for most samples. In contrast, when plotting the total difference against II, it becomes apparent that several samples emerge with different signs.

Hence, variance component I, concerning the reduced difference between a person-level and an integrated model, dominates the sign of the total difference, and thus gives a hint of whether the variance of the person-level or of the integrated GREG estimator is larger. Figure C.1 in Appendix C depicts a similar picture for case a). The only difference is that the total difference is always positive. Figure 5.14 plots the average household size within every MC sample against

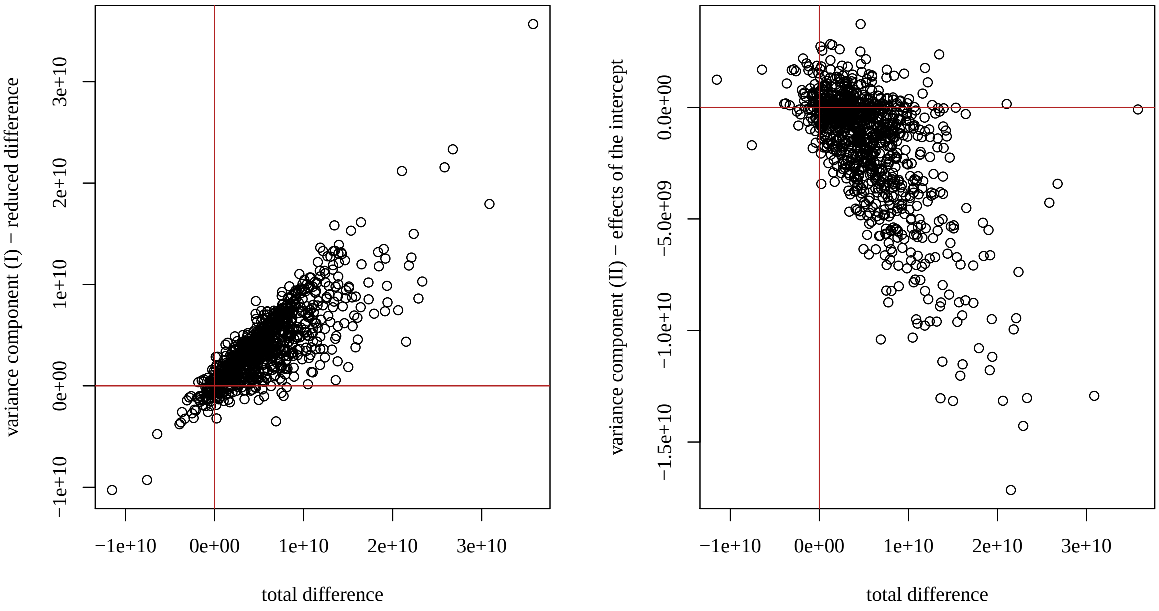

Figure 5.13: Plots of the total difference against variance components I and II for case b) and $m = 1500$

variance components I and II. It becomes obvious that both terms are unrelated to the average household size. The result can be explained through the fact that under normality and simple random sampling, the sample distributions of the point and variance estimator are independent. The average household sample size is not a point estimate, but a fixed value. The same is true for Figure C.2 in Appendix C. Figure 5.15 plots the variance components against each other. Positive total differences are indicated by circles, negative total differences by triangles. The amount of the total difference is, moreover, highlighted by the color. Blue refers to a small total difference, red to a large total difference. It becomes apparent that the higher variance component I, the higher the total difference. In contrast, the higher variance component II, the smaller the total difference. Therefore, if variance component I - describing the difference in the variances of a naïve and a reduced household-level GREG estimator - is small, it is more prevalent that variance of an integrated household-level GREG estimator exceeds the variance of a naïve GREG estimator. The following two sections separately analyze variance components I and II in more detail.

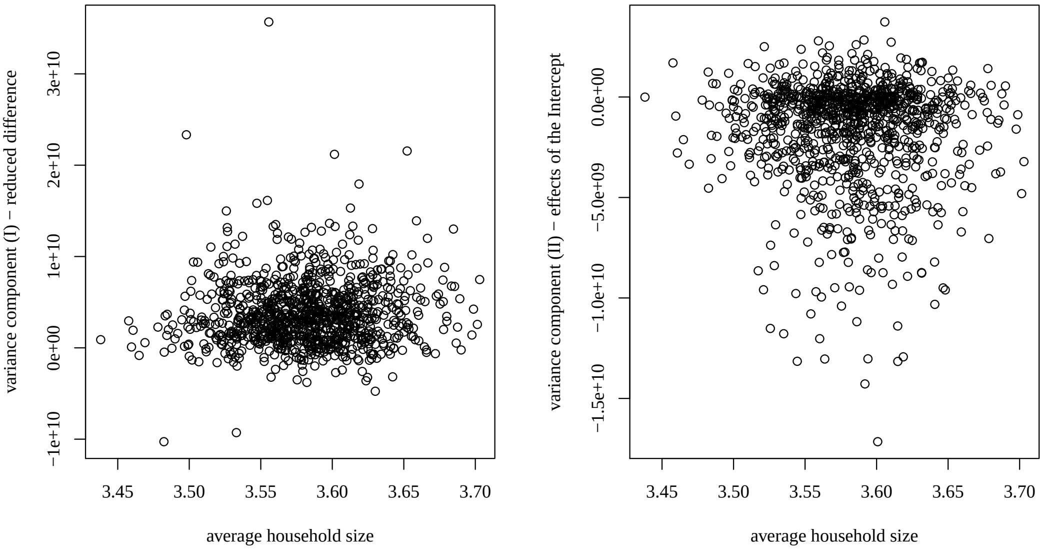

Figure 5.14: Plots of the intercept and reduced difference against the average household size for case b) and $m = 1500$

5.2.5.1 Variance Component I - Reduced Difference of the Same Dimension

To check the presumption that variance component I depends on the correlation between $\boldsymbol{x_i}$ and $\bar{\boldsymbol{x}}_i$, denoted as Cor$(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_i)$, we compute $\boldsymbol{D}_{\bar{x}}$ and $(\boldsymbol{B}_x - \boldsymbol{B}_{\bar{x}})$ in the cases of low and high correlation. For a better visualization via plots, we include only two auxiliaries concerning the same variable: once in its original form, x_i, and once as household mean value, $\bar{x}_i$. As auxiliary variables, we choose `age1` and `age4`. The former variable is characterized by a lower correlation between x_i and $\bar{x}_i$, Cor$(x_i, \bar{x}_i) = 0.51$. The latter variable is characterized by a higher correlation, Cor$(x_i, \bar{x}_i) = 0.69$. Since x_i and $\bar{x}_i$ refer to the same variable, Cor$(x_i, \bar{x}_i) = 0.51$ is the lowest correlation we found in our data set. Figure 5.16 plots the coefficients $D_{\bar{x}}$ and $(B_x - B_{\bar{x}})$ against variance component I. Notice that the scales of the y-axes in the upper and lower plots are equal. As expected, the point cloud in the upper left plot is higher located than in the upper right plot, which confirms that a higher correlation between x_i and $\bar{x}_i$ negatively affects $D_{\bar{x}}$. Furthermore, we observe that $D_{\bar{x}}$ and variance component I are positively related. Also the relation between the correlation between x_i and $\bar{x}_i$ and $(B_x - B_{\bar{x}})$ is in line with our presumption explored in the previous section. Nevertheless, $(B_x - B_{\bar{x}})$ has no effect on variance component I. Therefore, it seems that variance component I is mainly driven by the coefficient $\boldsymbol{D}_{\bar{x}}$, describing the effect of $\bar{\boldsymbol{x}}_i$ on y_i when controlling for the effect of $\boldsymbol{x_i}$. Another question is whether we can deduce which variance dominated the total difference from the relation between the residuals. For this purpose, Figure C.3 in Appendix C plots the residuals of a person-level and a reduced household-level GREG estimator in cases of a positive and a negative variance component I. The residuals on the person-level model are aggregated per household, since its aggregated form enters into the variance formula and otherwise the person- and household-level

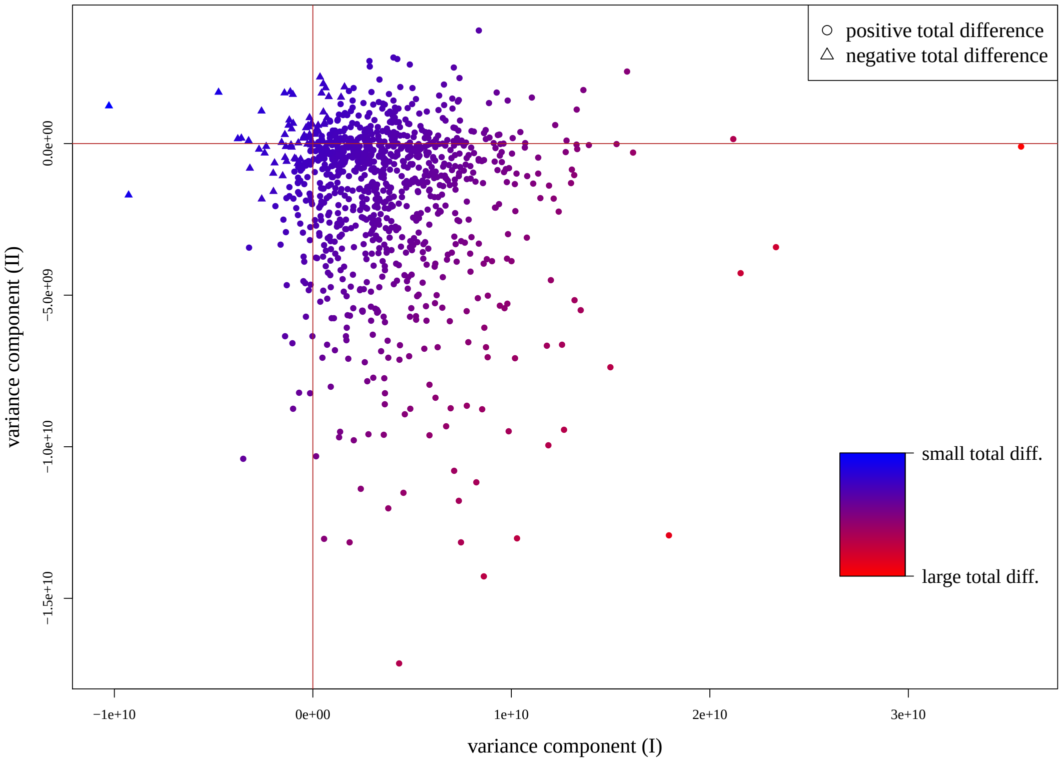

Figure 5.15: Plot of variance component I against variance component II for case b) and $m = 1500$

residuals cannot be presented in one plot. No differences are achieved between the left- and the right-hand plots. Only slight differences can be observed between a high or low correlation.

In Section 3.2, we argued that the integrated GREG estimator considers only the between-variance of the auxiliaries due to the replacement of the original auxiliary information by the constructed household mean values. Thus, Figure C.4 in Appendix C plots the within variance against variance component I. Surprisingly, the within variance seems to be unrelated from variance component I. A possible reason is that the effect of the within variance is superimposed by other effects.

5.2.5.2 Variance Component II - Effects of the Intercept

As mentioned in Section 5.2.4.2, the prediction of which factors affect variance component II is critical, since the relevance of the intercept is of a more technical nature. We know only that $\boldsymbol{F_x}$ depends on the auxiliaries. Even if Figure 5.17 approves that $\boldsymbol{F_x}$ is slightly higher for a lower correlation, there seems to be no effect of $\boldsymbol{F_x}$ on variance component II. Hence, the amount of variance component II, capturing the effect of the intercept, is difficult to predict.

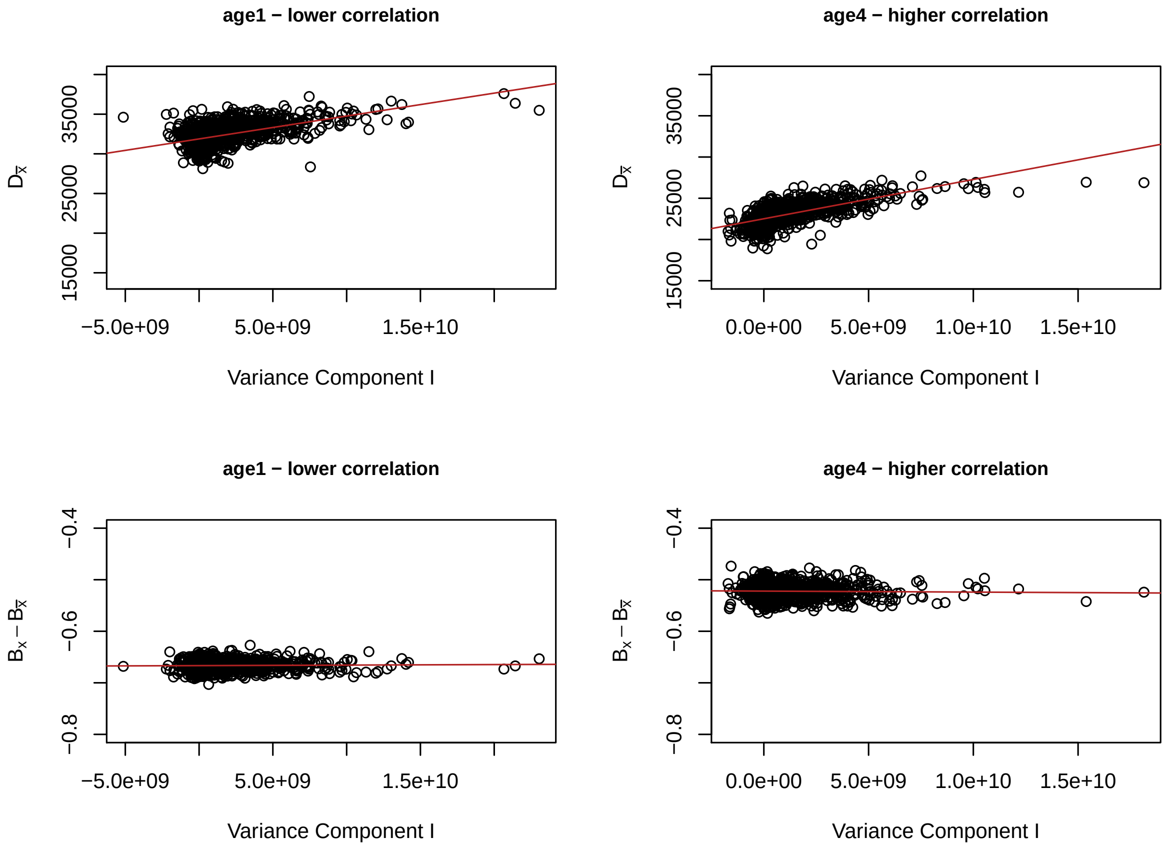

Figure 5.16: Plots of $D_{\bar{x}}$ and $(B_x - B_{\bar{x}})$ against variance component I for case b) and $m = 1500$

In conclusion, objective function (5.64), specifying the correct efficiency comparison between a person-level and an integrated GREG estimator, is determined by variance components I and II. Variance component I is mainly driven by $\boldsymbol{D}_{\bar{x}}$ from the overlap model (5.59). We found that the lower is the correlation between the original auxiliary variable $\boldsymbol{x}_i$ and its household mean value, the higher $\boldsymbol{D}_{\bar{x}}$. Then, the higher $\boldsymbol{D}_{\bar{x}}$, the higher variance component I, which describes the difference between the variances of a person-level and a reduced household-level model. Variance component I, describing the effect of the intercept, which determines the difference in dimension between a person-level and an integrated GREG estimator, is difficult to predict.

5.3 Further Application Field for the Decomposition of the Coefficients

Another promising application field for the decomposition of coefficients presented in Section 5.2.1.1 is econometrics. Econometrics applies statistical methods to empirical data to evaluate and develop econometric theory (cf. Greene, 2003, p. 1; Wooldridge, 2013, p. 2). A widespread

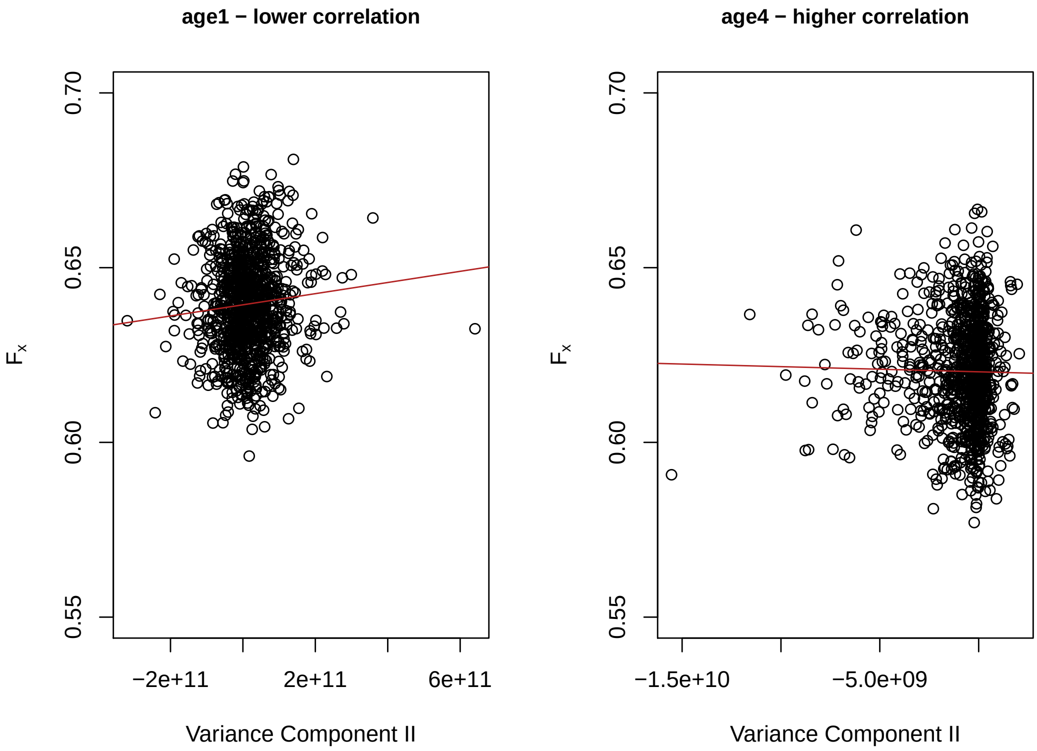

Figure 5.17: Plots of F_x against variance component II for case b) and $m = 1500$

econometric method is linear regression analysis, where the relationship between a variable of interest and certain explanatory variables is modeled by a linear function. Often, various explanatory variables are available from one (or more) data sources. This gives rise to the question of which explanatory variables should be included into the model. Even if the foundation of a model should always be built on economic theory or preliminary studies, several competitive models containing different sets of explanatory variables can be chosen. The potential *best* model can be assessed via goodness of fit criteria, which indicate how well a model fits the observed data. A well-known goodness of fit criterion is the coefficient of determination, abbreviated by R^2. Further measures are Mallow's complexity parameter, the Akaike information criterion, the Bayes information criterion or cross-validation. The interested reader is referred to for example Fahrmeir et al. (2007, p. 162) for more details on goodness of fit criteria.

The coefficient of determination R^2 evaluates the explanatory power of a linear regression model via the ratio of the explained variation (regression sum of squares abbreviated by SSR) and the total variation (total sum of squares abbreviated by SST) of a variable of interest

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i \in U_p} (\hat{y}_i - \bar{y})^2}{\sum_{i \in U_p} (y_i - \bar{y})^2}$$

with y_i as observed values of the variables of interest, $\hat{y}_i$ as fitted values predicted by the linear model, and $\bar{y}$ as mean value (cf. Backhaus et al., 2008, p. 72). The value of R^2 ranges from

0 to 1. It can be interpreted as percentage explanation of the variation of a variable of interest by the explanatory variables in the model. Thus, higher values of R^2 indicate a good fit of the predicted values $\hat{y}_i$ (cf. Schlittgen, 2008, p. 420).

A question arising in this context regards how R^2 is affected when additional explanatory variables are added to the initial model. The question is answered by the difference between the coefficients of determination of two nested models

$$R^2_{II} - R^2_I.$$

Subscript I refers to the first model, whereas subscript II refers to the second model, which is nested within model I.

Another field of application for $R^2_{II} - R^2_I$ is survey statistics. Even if the GREG estimator is model-assisted, and thus its unbiasedness is unaffected by the correctness of the model, its efficiency depends on the explanatory power of the assisting model. The difference $R^2_{II} - R^2_I$ delivers a criterion to decide which auxiliary variables (with known totals) should be included into the assisting model to increase the efficiency of the estimator.

The decomposition of coefficients presented above is helpful for explaining the difference $R^2_{II} - R^2_I$. The idea is to use the overlap of two nested models to relate their coefficients. Inserting the decomposition into the difference $R^2_{II} - R^2_I$ reveals a deeper understanding of which factors influence the supplementary degree of explanation driven by including additional explanatory variables.

As before, we initially focus on simple models containing two variables in order to visualize the decomposition via graphs. Subsequently, we extend our findings to multiple explanatory variables. Suppose that model I contains x_i as initial explanatory variable and is therefore given by

$$y_{Ii} = B_x x_i + r_i^B.$$

Beyond x_i, model II contains z_i as an additional explanatory variable and is expressed as

$$y_{IIi} = D_x x_i + D_z z_i + r_i^D.$$

Model I is nested within model II. Hence, the difference $R^2_I - R^2_{II}$ quantifies the supplementary explained variation of y_i when adding z_i as additional explanatory variable besides x_i into the model. To apply the mediation model, we interpret the additional explanatory variable z_i as the mediator variable and B_x obtained from model I as the total effect. Then, the total effect B_x can be decomposed into a direct effect of x_i on y_i controlling for z_i plus an indirect effect of x_i via z_i (illustrated by Figure 5.18). To quantify the indirect effect, we specify the following auxiliary model

$$z_i = F_x x_i + r_i^{F_x}.$$

$$\begin{tikzcd} & z_i \arrow[dr, "D_z"] & \\ x_i \arrow[ur, "F_x"] \arrow[rr, "D_x"'] & & y_i \end{tikzcd}$$

Figure 5.18: Mediation model applied to two nested models

Consequently, the decomposition of the total effect of the initial explanatory variable x_i is given by

$$B_x = D_x - F_x D_z.$$

Now, we translate the decomposition of B_x to the case of multiple explanatory variables. Let $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{iq}, \ldots, x_{iQ})^T$ and $\boldsymbol{z_i} = (z_{i1}, \ldots, z_{ik}, \ldots, z_{iK})^T$ be two vectors of dimensions Q and K of explanatory variables for individual i. Consider the following two multiple models, whereby model II is nested within model I,

$$\begin{aligned} y_{Ii} &= \boldsymbol{B_x}^T \boldsymbol{x_i} + r_i^B \\ y_{IIi} &= \boldsymbol{D_x}^T \boldsymbol{x_i} + \boldsymbol{D_z}^T \boldsymbol{z_i} + r_i^D, \end{aligned}$$

with $\boldsymbol{D_x} = (D_{x_1}, \ldots, D_{x_Q})^T$ and $\boldsymbol{D_z} = (D_{z_1}, \ldots, D_{z_K})^T$. Both models have the initial explanatory variables $\boldsymbol{x_i}$ in common. Beyond $\boldsymbol{x_i}$, model II contains the additional explanatory variables $\boldsymbol{z_i}$. Due to multidimensionality, we have to specify for every k-th additional explanatory variable z_{ik} with $k = 1, \ldots, K$ an auxiliary model

$$\begin{aligned} z_{i1} &= \boldsymbol{F_x^1}^T \boldsymbol{x_i} + r_i^{F_x^1} \\ \vdots \quad & \quad \vdots \\ z_{ik} &= \boldsymbol{F_x^k}^T \boldsymbol{x_i} + r_i^{F_x^k} \\ \vdots \quad & \quad \vdots \\ z_{iK} &= \boldsymbol{F_x^K}^T \boldsymbol{x_i} + r_i^{F_x^K} \end{aligned}$$

with $\boldsymbol{F_x^k} = (F_{x_1}^k, \ldots, F_{x_Q}^k)^T$. Thus, in each regression, the k-th additional explanatory variable z_{ik} is regressed on the complete set of initial explanatory variables $\boldsymbol{x_i}$, which is common to both models. Suppose all K vectors $\boldsymbol{F_x^k}$ are combined to one matrix $\boldsymbol{F_x} = (\boldsymbol{F_x^1}, \ldots, \boldsymbol{F_x^k}, \ldots, \boldsymbol{F_x^K})$

of the form

$$\boldsymbol{F_x} = \begin{pmatrix} F^1_{x_1} & \dots & F^k_{x_1} & \dots & F^K_{x_1} \\ \vdots & & \vdots & & \vdots \\ F^1_{x_q} & \dots & F^k_{x_q} & \dots & F^K_{x_q} \\ \vdots & & \vdots & & \vdots \\ F^1_{x_Q} & \dots & F^k_{x_Q} & \dots & F^K_{x_Q} \end{pmatrix}.$$

Then, analogous to the decomposition in Lemma 1 the multiple total effect $\boldsymbol{B_x}$ can be decomposed into

$$\underset{Q\times 1}{\boldsymbol{D_x}} = \underset{Q\times 1}{\boldsymbol{B_x}} - \underset{Q\times K}{\boldsymbol{F_x}} \cdot \underset{K\times 1}{\boldsymbol{D_z}}. \tag{5.65}$$

For a better understanding of $\boldsymbol{F_x} \cdot \boldsymbol{D_z}$ in equation (5.65), consider the q-th row given by $(F^1_{x_q} \cdot D_{z_1} + \ldots + F^k_{x_q} \cdot D_{z_k} + \ldots + F^K_{x_q} \cdot D_{z_K})$. Thus, by construction of $\boldsymbol{F_x}$ the multiplication $\boldsymbol{F_x} \cdot \boldsymbol{D_z}$ causes each auxiliary model the k-th coefficient $F^k_{x_q}$ to be multiplied by D_{z_k}, which is obtained from model II with respect to the same k-th explanatory variable.

Given the multiple decomposition of $\boldsymbol{B_x}$, the following result explains how $R^2_{II} - R^2_I$ is affected when additional explanatory variables are added to the initial model.

Result 11. ***Difference $R^2_{II} - R^2_I$ of Two Nested Models***
Consider two predicted values $\hat{y}_{Ii} = \boldsymbol{B_x}^T \boldsymbol{x_i}$ and $\hat{y}_{IIi} = \boldsymbol{D_x}^T \boldsymbol{x_i} + \boldsymbol{D_z}^T \boldsymbol{z_i}$, whereby model I is nested within model II. Then the following applies

$$R^2_{II} - R^2_I = \frac{\sum_{i\in U_p} (\boldsymbol{D_z}^T \boldsymbol{r_i^{F_x}})^2}{\sum_{i\in U_p} (y_i - \bar{y})^2}.$$

The residual vector $\boldsymbol{r_i^{F_x}} = (r_i^{F_x^1}, \ldots, r_i^{F_x^k}, \ldots, r_i^{F_x^K})^T$ contains the residuals from all K auxiliary models $z_{ik} = \boldsymbol{F_x^k}^T \boldsymbol{x_i} + r_i^{F_x^k}$ with $k = 1, \ldots, K$.

Proof. The difference of the coefficients of determination of two nested models is given by

$$\begin{aligned} R^2_{II} - R^2_I &= \frac{\sum_{i\in U_p} (\hat{y}_{iII} - \bar{y})^2}{\sum_{i\in U_p} (y_i - \bar{y})^2} - \frac{\sum_{i\in U_p} (\hat{y}_{iI} - \bar{y})^2}{\sum_{i\in U_p} (y_i - \bar{y})^2} \\ &= \frac{\sum_{i\in U_p} (\hat{y}^2_{iII} - 2\hat{y}_{iII}\bar{y} + \bar{y}^2) - \sum_{i\in U_p} (\hat{y}^2_{iI} - 2\hat{y}_{iI}\bar{y} + \bar{y}^2)}{\sum_{i\in U_p} (y_i - \bar{y})^2} \\ &= \frac{\sum_{i\in U_p} (\hat{y}^2_{iII} - \hat{y}^2_{iI}) - 2\bar{y} \sum_{i\in U_p} (\hat{y}_{iII} - \hat{y}_{iI})}{\sum_{i\in U_p} (y_i - \bar{y})^2}. \end{aligned} \tag{5.66}$$

The first term in the numerator in (5.66) can be rewritten as

$$\begin{aligned}
&\sum_{i \in U_p} (\hat{y}_{iII}^2 - \hat{y}_{iI}^2) \\
&\quad = \sum_{i \in U_p} (\hat{y}_{iII} + \hat{y}_{iI})(\hat{y}_{iII} - \hat{y}_{iI}) \\
&\quad = \sum_{i \in U_p} (\boldsymbol{D_x}^T \boldsymbol{x_i} + \boldsymbol{D_z}^T \boldsymbol{z_i} + \boldsymbol{B_x}^T \boldsymbol{x_i})(\boldsymbol{D_x}^T \boldsymbol{x_i} + \boldsymbol{D_z}^T \boldsymbol{z_i} - \boldsymbol{B_x}^T \boldsymbol{x_i}).
\end{aligned}$$

Inserting the multiple decomposition $\boldsymbol{D_x} = \boldsymbol{B_x} - \boldsymbol{F_x}\boldsymbol{D_z}$ from (5.65) yields

$$\begin{aligned}
&= \sum_{i \in U_p} \left((\boldsymbol{B_x} - \boldsymbol{F_x}\boldsymbol{D_z})^T \boldsymbol{x_i} + \boldsymbol{D_z}^T \boldsymbol{z_i} + \boldsymbol{B_x}^T \boldsymbol{x_i}\right)\left((\boldsymbol{B_x} - \boldsymbol{F_x}\boldsymbol{D_z})^T \boldsymbol{x_i} + \boldsymbol{D_z}^T \boldsymbol{z_i} - \boldsymbol{B_x}^T \boldsymbol{x_i}\right) \\
&= \sum_{i \in U_p} \left(2\boldsymbol{B_x}^T \boldsymbol{x_i} - \boldsymbol{D_z}^T (\boldsymbol{F_x}^T \boldsymbol{x_i} - \boldsymbol{z_i})\right)\left(-\boldsymbol{D_z}^T (\boldsymbol{F_x}^T \boldsymbol{x_i} - \boldsymbol{z_i})\right). \qquad (5.67)
\end{aligned}$$

Since $\boldsymbol{z_i} - \boldsymbol{F_x}^T \boldsymbol{x_i}$ determines the residual vector $\boldsymbol{r_i^{F_x}} = (r_i^{F_{x_1}^1}, \dots, r_i^{F_x^k}, \dots, r_i^{F_x^K})^T$, (5.67) can be rewritten as

$$\begin{aligned}
&= -\,2 \sum_{i \in U_p} \boldsymbol{B_x}^T \boldsymbol{x_{1i}} \boldsymbol{D_z}^T \boldsymbol{r_i^{F_x}} + \sum_{i \in U_p} (\boldsymbol{D_z}^T \boldsymbol{r_i^{F_x}})^2 \\
&= -\,2 \underbrace{\sum_{i \in U_p} \boldsymbol{D_z}^T \boldsymbol{r_i^{F_x}} \boldsymbol{x_{1i}}^T \boldsymbol{B_x}}_{=0} + \sum_{i \in U_p} (\boldsymbol{D_z}^T \boldsymbol{r_i^{F_x}})^2. \qquad (5.68)
\end{aligned}$$

The first term in (5.68) can be rearranged to

$$\sum_{i \in U_p} \boldsymbol{D_z}^T \boldsymbol{r_i^{F_x}} \boldsymbol{x_i}^T \boldsymbol{B_x} = \boldsymbol{D_z}^T (\sum_{i \in U_p} \boldsymbol{r_i^{F_x}} \boldsymbol{x_i}^T) \boldsymbol{B_x}.$$

From the least squares theory (cf. Greene, 2003, Section 6.4; Wooldridge, 2013, Section 3.2), we know that the residuals from the k-th auxiliary model sum up to zero, if the model contains an intercept (see Section 5.2.1.2), i.e. $\sum_{i \in U_p} r_i^{F_x^k} = 0$. From $\sum_{i \in U_p} r_i^{F_x^k} x_{iq} = 0$ for all $k = 1, \dots, K$ and $q = 1, \dots, Q$, it follows that $\sum_{i \in U_p} \boldsymbol{r_i^{F_x}} \boldsymbol{x_i}^T = \boldsymbol{0}$, where in this case $\sum_{i \in U_p}$ is defined as a component-wise summation within the matrix. Consequently, it is valid that $\sum_{i \in U_p} \boldsymbol{D_z}^T \boldsymbol{r_i^{F_x}} \boldsymbol{x_i}^T \boldsymbol{B_x} = 0$.

We continue with the second term in the numerator of (5.66). After some algebraic transformations, we obtain

$$\begin{aligned}
&2\bar{y}\sum_{i\in U_p}(\hat{y}_{iII}-\hat{y}_{iI})\\
=&2\bar{y}\Big(\sum_{i\in U_p}\Big(\boldsymbol{D_x}^T\boldsymbol{x_i}+\boldsymbol{D_z}^T\boldsymbol{x_z}-\sum_{i\in U_p}\boldsymbol{B_x}^T\boldsymbol{x_i}\Big)\Big)\\
=&2\bar{y}\Big(\sum_{i\in U_p}\Big((\boldsymbol{B_x}-\boldsymbol{F_x}\boldsymbol{D_z})^T\boldsymbol{x_i}+\boldsymbol{D_z}^T\boldsymbol{z_i}\Big)-\sum_{i\in U_p}\boldsymbol{B_x}^T\boldsymbol{x_i}\Big)\\
=&2\bar{y}\Big(\sum_{i\in U_p}\boldsymbol{B_x}^T\boldsymbol{x_i}-\sum_{i\in U_p}\boldsymbol{D_z}^T\underbrace{(\boldsymbol{z_i}-\boldsymbol{F_x}^T\boldsymbol{x_i})}_{\boldsymbol{r_i^{F_x}}}-\sum_{i\in U_p}\boldsymbol{B_x}^T\boldsymbol{x_i}\Big)\\
=&-2\bar{y}\sum_{i\in U_p}\boldsymbol{D_z}^T\boldsymbol{r_i^{F_x}}\\
=&0.
\end{aligned}$$

Finally, inserting (5.68) into (5.66) yields

$$R_1^2-R_2^2=\frac{\sum_{i\in U_p}(\boldsymbol{D_z}^T\boldsymbol{r_i^{F_x}})^2}{\sum_{i\in U_p}(y_i-\bar{y})^2}.$$

Therefore, Result 11 is proven. □

We learn from Result 11 that $R_I^2-R_{II}^2$ depends on

a) $\boldsymbol{D_z}$ describing the effect of the additional explanatory variables $\boldsymbol{z_i}$ on y_i controlling for the effects of the initial explanatory variables $\boldsymbol{x_i}$ already included in the model, and

b) $\boldsymbol{r_i^{F_x}}$ describing the remaining variation of the additional explanatory variables $\boldsymbol{z_i}$ not explained by the initial explanatory variables $\boldsymbol{x_i}$.

In other words, the difference $R_I^2-R_{II}^2$ depends on the effect of the additional explanatory variables not explained by the initial explanatory variables and on the degree to which the initial explanatory variables help to explain the additional explanatory variables. Consequently, the higher the correlation between the initial explanatory variables and the additional explanatory variables, the lower the increased degree of explanation of the variation of the variable of interest. With respect to the Venn diagrams, extensively introduced in Section 5.1.2.2, this result is not surprising. The higher the correlation between the initial and the additional explanatory variables, the higher the common variation not used to explain the variable of interest. In a Venn diagram, the common variation is illustrated by the intersections of the circles of the initial and the additional explanatory variables. In this regard, Result 11 is helpful within the process of variable selection. It affords a deeper understanding of how the implementation of additional explanatory variables causes supplementary explanatory power of the model. Further variable selection strategies can be found in Bethlehem et al. (2011, pp. 261-274) or Fahrmeir et al. (2007, pp. 152-180).

It is important to remark that R^2 always increases when additional variables are added to the model (cf. Wooldridge, 2013, p. 254). As a remedy, the adjusted coefficient of determination incorporates a penalty term to adjust for the number of explanatory variables and observations (cf. Fahrmeir et al., 2007, pp. 98-100). A detailed discussion of the adjusted coefficient of determination and further limitations of R^2 are given in Backhaus et al. (2008, pp. 72-76) or Schlittgen (2008, pp. 420-422).

A similar result for $R_I^2 - R_{II}^2$ as given in Result 11 can be found in Greene (2003, p. 254). Also Seber (1977) derived the difference between two coefficients of determination. However, the proof differs from our proof as it relies on geometric arguments of linear regression analysis.

To verify the correctness of Result 11, we run a simulation study based on the simulation setting defined in Section 3.4.1. The variables of interest and the auxiliary variables are known from Tables 3.6 and 3.7. The initial explanatory variable set $\boldsymbol{x}_i$ contains `sex` and `ms` (six indicator variables). The additional explanatory variable set $\boldsymbol{z}_i$ consists of `age` (four indicator variables). To determine the *true value* of $R_I^2 - R_{II}^2$ we use the R command lm(). The calculations are based on one MC sample ($r = 1$). Table 5.6 validates the correctness of Result 11, since there is no difference between the left and the right columns.

Table 5.6: Difference of $R_{II}^2 - R_I^2$ for various variables of interest and for $r = 1$

	$R_{II}^2 - R_I^2$	$\frac{\sum_{i\in U_p}(\boldsymbol{D}_{\boldsymbol{x}_2}{}^T\boldsymbol{r}_i^{\boldsymbol{F}_{\boldsymbol{x}_1}})^2}{\sum_{i\in U_p}(y_i-\bar{y})^2}$
inc	0.0207	0.0207
soc	0.0071	0.0071
sel	0.0083	0.0083
act1	0.0539	0.0539
act2	0.0099	0.0099
act3	0.0781	0.0781
inc_hs1	0.0118	0.0118
inc_hs2	0.0102	0.0102
inc_hs3	0.0042	0.0042
inc_hs4	0.0017	0.0017
inc_hs5	0.0012	0.0012
inc_hs6	0.0001	0.0001
bene_age1	0.0084	0.0084
bene_age2	0.0159	0.0159
bene_age3	0.0167	0.0167
bene_age4	0.0083	0.0083

5.4 Summary and Conclusion

The objective of this chapter was to derive an efficiency comparison of a person-level and an integrated household-level GREG estimator. Initially, we presented the efficiency comparison of Steel and Clark (2007) and discussed two issues. At first, they tacitly assumed that the auxiliaries of a person-level GREG estimator sum up per household to the auxiliaries of an integrated household-level GREG estimator. However, we showed that the per-household summation of the person-level information results in a household-level auxiliary vector without an intercept. In a MC simulation study, we clarified that, through the exclusion of the intercept, Steel and Clark (2007) underestimated the correct variance of an integrated GREG estimator, in particular if large households prevail in the sample. Secondly, applying Venn diagrams and the FWL theorem, we demonstrated that although Steel and Clark (2007) justified the interpretation of their final result using the argument of *controlling for*, their approach considerably differs from the concept of *controlling for* originated from multiple regressions. The difference between both interpretation approaches is quantified by the common variation of the auxiliaries $\boldsymbol{x}_i$ and $\bar{\boldsymbol{x}}_i$ and depends on the household sizes.

Due to these issues, we derived in Section 5.2 an own efficiency comparison between a person-level and an integrated household-level GREG estimator. The proceeding to provide such an efficiency comparison was twofold. In a first step, we aim at separating the effect of the intercept from the variance of an integrated GREG estimator. The intercept constitutes the difference in dimension between a person-level and an integrated model. To solve this problem, we decomposed the variance of an integrated household-level GREG estimator into the variance of a reduced household-level GREG estimator and into a term that captures the effect of the intercept disregarded by the reduced household-level model. The model of the reduced household-level GREG estimator, excluding an intercept, is of the same dimension as the model of a person-level GREG estimator. The decomposition consists of three steps. We started with decomposing the coefficients of an integrated household-level model (Lemma 1) by applying mediation models known from psychology and sociology. The decomposition of the integrated coefficient is impressive, as in multiple regressions the coefficients are calculated *ceteris paribus* and therefore incorporate the covariances of the auxiliary variables. We continued with translating the decomposition of the integrated coefficient to the corresponding residuals (Lemma 2). For this purpose, we introduced artificially constructed pseudo-residuals, which permit us to exactly quantify the effect of the intercept on the variance disregarded by the reduced household-level model. Finally, we extended our findings to the decomposition of the sum of squared residuals (Result 8). The decomposition is powerful, as even if the power of two is a non-linear transformation, the sum of the squared residuals of the original integrated household-level model equals the sum of squared residuals for the separated regressions as it would be with a linear transformation. Thus, when deriving the decomposition of the variance, we can skip all mixed terms emerging when multiplying out the binomial formula. In consequence, with the decomposition of the variance of the integrated GREG estimator, we are capable of deriving a correct efficiency comparison between a person-level and an integrated GREG estimator.

For the decomposition of the variance of the integrated GREG estimator, we define coefficients from models without an intercept, pseudo-residuals, and separating residuals. In order to elaborate the differences between the coefficients and residuals obtained from originate regression models to our defined coefficients and residuals, we deduced their properties.

In a second step, we derived a functional relationship between the $\boldsymbol{B}_p$ and $\boldsymbol{B}_h$ as coefficients from the person-level and the reduced household-level GREG estimator. The fact that both coefficients are computed at different levels hampers the pursuit of discovering a relationship between them. As a remedy, we exploited that the reduced household-level coefficient can either be computed at household-level, using OLS, or at person-level, using GLS. Hence, we considered two person-level models to derive a relationship between $\boldsymbol{B}_p$ and $\boldsymbol{B}_h$. Then, we related both coefficients by constructing an overlap model, which simultaneously contains the auxiliaries of both coefficients. With the overlap model as a common starting point, we decomposed $\boldsymbol{B}_p$ and $\boldsymbol{B}_h$ by applying two mediation models into the same coefficients obtained from the overlap model. However, writing $\boldsymbol{B}_p$ as function of $\boldsymbol{B}_h$ is not straightforward, because the arrays of the decomposed coefficients differ. To handle this obstacle, we made use of the form of the coefficient resulting from regressing the original auxiliary on its constructed household mean values derived in Lemma 4 (or in Lemma 3 for simple models). However, for dimensions $Q > 2$, the proof becomes computationally cumbersome. Since the problem is the analytical representation of the inverse of $\sum_{i \in U_p} \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^T$, we proposed reducing the dimension of the matrix on which the inverse is applied without the reduction of Q itself. For this purpose, we applied the concept of partial regressions and the FWL theorem. Finally, Result 10 (or Result 9 for simple models) allowed us to write $\boldsymbol{B}_p$ as function of $\boldsymbol{B}_h$ and certain coefficients required in the decomposition.

To conclude, with Results 8 and 10 (or 9) we are capable to provide an efficiency comparison between the variances of a person-level and an integrated GREG estimator. Accordingly, the difference is determined by variance component I, describing the difference of the variances of a reduced household-level and a person-level GREG estimator, and variance component II, which captures the effects of the intercept on the variance of an integrated household-level model but is disregarded by the reduced household-level model. Variance component I depends on the correlation between the original auxiliaries and the constructed household mean values. The effect of variance component I is difficult to estimate since it depends on the intercept.

It should be noted that the interpretation of the presented efficiency comparison has some limitations. It offers only an answer to the following question: How is the efficiency of an integrated GREG estimator affected by the consistency requirement compared with a person-level GREG estimator? However, it does not answer the question: Is integrated weighting preferable to a person-level GREG estimator? The reason for this is that both estimators under consideration pursue different targets. If the objective is to ensure consistent estimates between the person and household level, integrated weighting can be applied but not a person-level GREG estimator. However, if consistent estimates are required, we suggest using our proposed modified extended GREG estimator, introduced in Chapter 4 instead of integrated weighting. If, on the other hand, consistency is not required, integrated weighting is never the preferable choice because it utilizes constructed household mean values instead of the original person-level information and enforces equal weights for all household members.

In Section 5.3, we discussed a further application field for the introduced decomposition of the coefficients within the context of variable selection in econometrics and survey statistics. Hence, we analyzed the difference of two coefficients of determination R^2 to answer the question how it is affected when additional explanatory variables are added to the initial model. We learn from Result 11 that the difference of two coefficients of determination depends on the effect of the additional explanatory variables not explained by the initial explanatory variables and the degree to which the initial explanatory variables explain the additional explanatory variables.

6 The Variance Formula of GREG Estimators under Cluster Sampling and the Proposed Hybrid GREG Estimator

In household surveys, information is collected on both the persons and households. Thus, the assisting model of the GREG estimator can be established at either the person or the household level. However, the variance formulas of both GREG estimators depend on the per household aggregated variables (cf. Särndal et al., 1992, p. 307). The initial level of modeling is, therefore, disregarded in the variance formula. This can be interpreted that in the variance formula the households, or in general the clusters, are treated as the ultimate sampling unit which corresponds to fundamental surveys textbooks:

- Lohr (2009, p. 171): "No new ideas are introduced to carry out one-stage cluster sampling; we simply use the results for simple random sampling with the PSU totals as the observations."

- Thompson (2002, p. 129): "[...] one could dispense with the concept of the secondary units, regarding the primary units as the sampling units and using, as the variable of interest of any primary unit, the total of the y-values of the secondary units within it."

The aim of this chapter is twofold: First, we study the consequences of the aggregated form of the variance formula under cluster sampling for person-level GREG estimators. One consequence is that there is a mismatch between the residuals in the minimization problem, which deliver the coefficient of the point estimator, and the residuals in the variance formula. Another consequence is that the optimal estimator, which is the estimator that minimizes the variance, is based on the per household aggregated variables. This implies that under cluster sampling one should always use the aggregated person-level information even if the variable of interest is a person characteristic. We elaborate that this implication is particularly critical for large and heterogeneous households. Second, as a remedy, we develop a hybrid GREG estimator that compromises between an optimal and a person-level GREG estimator.

The remainder of this chapter is organized as follows: Section 6.1 extensively discusses the variance formula of GREG estimators under cluster sampling. Section 6.2 reviews the literature on alternative variance formulas. In Section 6.3, we develop the hybrid GREG estimator that balances between optimality and person-level modeling. A simulation study verifies the theoretical discussed consequences for person-level GREG estimators and validates the performance of the proposed hybrid GREG estimator (Section 6.4). Section 6.5 summarizes the results.

6.1 Consequences of the Variance Formula on Person-Level GREG Estimators

In this section, we recapitulate the variance formula of GREG estimators under cluster sampling given for example in Särndal et al. (cf. 1992, p. 307).[1] As a short reminder, the population of the households is given by $U_h = \{1, \dots, M\}$. The population of persons in a certain household g is described by U_g. The number of persons within a household g is denoted by N_g. The auxiliary vector of person i is determined by $\boldsymbol{x_i} = (x_{i1}, \dots, x_{iQ})^T = (1, x_{i2}, \dots, x_{iQ})^T$. The per-household aggregated person-level auxiliary vector is given by $\boldsymbol{x_g} = \sum_{i \in U_g} \boldsymbol{x_i} = (x_{g1}, \dots, x_{gQ})^T = (N_g, x_{g2}, \dots, x_{gQ})^T$. It should be distinguished from the household-level auxiliary vector $\boldsymbol{a_g}$, as $\boldsymbol{x_g}$ contains the number of persons within the household, N_g, instead of an intercept. The variable of interest of person i is denoted by y_i. We assume single-stage cluster sampling, which means all persons within a selected household are sampled. For the sake of convenience, we use the terms single-stage cluster sampling and cluster sampling synonymously. Note that $\triangle_{gk} = \pi_{gk} - \pi_g \pi_k$, with π_g as first-order inclusion probability of household g and π_{gk} as second-order inclusion probability of households g and k.

Table 6.1: Point estimator and its variances of a person-level GREG estimator under cluster sampling I

	Person-level GREG estimator
Assisting model ξ	$y_i = \boldsymbol{x_i}^T \boldsymbol{\beta_p} + \epsilon_i$
Point estimator	$\hat{T}_{y_p}^{\text{GREG}} = \hat{T}_{y_p}^{\text{HT}} + \boldsymbol{B_p}^T (\boldsymbol{T_x} - \boldsymbol{\hat{T}_x^{\text{HT}}})$
Variance	$V(\hat{T}_{y_p}^{\text{GREG}}) = \sum\limits_{g \in U_h} \sum\limits_{k \in U_h} \triangle_{gk} \frac{r_g^{B_p}}{\pi_g} \frac{r_k^{B_p}}{\pi_k}$ with $r_g^{B_p} = \sum\limits_{i \in U_g} (y_i - \boldsymbol{x_i}^T \boldsymbol{B_p}) = y_g - \boldsymbol{x_g}^T \boldsymbol{B_p}$

Table 6.1 summarizes the formulas of the point estimators and its variances. We complete this table in the following. It becomes apparent from the table that the point estimator is unaffected by the cluster sampling design compared to a simple random sampling. However, the variance of the person-level GREG estimator depends on the per-household aggregates of the variable of interest y_g and the auxiliary variables $\boldsymbol{x_g}$, although the assisting model ξ refers to the person level counterparts y_i and $\boldsymbol{x_i}$. The following two sections elaborate the consequences of the aggregated form of the variance for person-level GREG estimators.

[1] Initially, (Särndal et al., 1992, p. 307) derived the formulas under two-stage cluster sampling. For reasons of simplification, we generalize their formulas to single-stage cluster sampling. For this purpose, we skip the term in the variance accounting for the randomness emerging within the selection process at the second stage. Nevertheless, our conclusion drawn in the following, are also valid for two-stage cluster sampling.

6.1.1 Mismatch between the Residuals in the Minimization Problem and in the Variance Formula

Following the least squares theory (cf. Greene, 2003, Section 6.4; Wooldridge, 2013, Section 3.2), we derive the coefficients of GREG estimators by minimizing the sum of squared residuals. The minimization problem and the corresponding coefficient are outlined in Table 6.2.

Table 6.2: Point estimator and its variances of a person-level GREG estimator under cluster sampling II

Person-level GREG estimator	
Assisting model ξ	$y_i = \boldsymbol{x_i}^T \boldsymbol{\beta_p} + \epsilon_i$
Point estimator	$\hat{T}_{y_p}^{\text{GREG}} = \hat{T}_{y_p}^{\text{HT}} + \boldsymbol{B_p}^T (\boldsymbol{T_x} - \boldsymbol{\hat{T}}_{\boldsymbol{x}}^{\text{HT}})$
Variance	$V(\hat{T}_{y_p}^{\text{GREG}}) = \sum\limits_{g \in U_h} \sum\limits_{k \in U_h} \triangle_{gk} \frac{r_g^{B_p}}{\pi_g} \frac{r_k^{B_p}}{\pi_k}$ with $r_g^{B_p} = \sum\limits_{i \in U_g} (y_i - \boldsymbol{x_i}^T \boldsymbol{B_p}) = y_g - \boldsymbol{x_g}^T \boldsymbol{B_p}$
Minimization problem	$\min\limits_{\boldsymbol{B_p}} \sum\limits_{g \in U_h} \sum\limits_{i \in U_g} \left(r_i^{B_p} \right)^2$ with $r_i^{B_p} = y_i - \boldsymbol{x_i}^T \boldsymbol{B_p}$
Resulting coefficient	$\boldsymbol{B_p} = (\sum\limits_{g \in U_h} \sum\limits_{i \in U_g} \boldsymbol{x_i} \boldsymbol{x_i}^T)^{-1} \sum\limits_{g \in U_h} \sum\limits_{i \in U_g} \boldsymbol{x_i} y_i$

We learn that the coefficient $\boldsymbol{B_p}$ is derived by minimizing the sum of squared person-level residuals $r_i^{B_p}$. However, these residuals in the minimization problem (fourth row) do not match to the per-household aggregated residuals $r_g^{B_p}$ used in the variance formula (third row). The reason for the mismatch is the order of the sum and the product. In consequence, the point estimator and its variance are not compatible.

To make the mismatch between the residuals in the minimization problem and in the variance formula more obvious, we temporarily assume that the households are selected by simple random sampling. Given the first- and second-order inclusion probabilities under simple single-stage cluster sampling

$$\pi_g = \frac{m}{M} \qquad \text{for} \quad g = \{1, \dots, M\}$$

$$\pi_{gk} = \frac{m}{M} \frac{(m-1)}{(M-1)} \qquad \text{for} \quad g, k = \{1, \dots, M\},\ g \neq k$$

the variance formula of the person-level GREG estimator simplifies to

$$\begin{aligned} V(\hat{T}_{y_p}^{GREG}) &= \sum_{g\in U_h}\sum_{k\in U_h} \triangle_{gk}\frac{r_g^{B_p}}{\pi_g}\frac{r_k^{B_p}}{\pi_k} \\ &= \sum_{g\in U_h}\triangle_{gg}\frac{{r_g^{B_p}}^2}{\pi_g} + \sum_{g\in U_h}\sum_{\substack{k\in U_h\\ k\neq g}}\triangle_{gk}\frac{r_g^{B_p}}{\pi_g}\frac{r_k^{B_p}}{\pi_k} \\ &= \frac{M^2}{m}\left(1-\frac{m}{M}\right)\frac{1}{m-1}\sum_{g\in U_h}{r_g^{B_p}}^2. \end{aligned}$$

Table 6.3 summarizes the results under simple cluster sampling where the households are drawn by simple random sampling. Note that the point estimators remain unchanged compared to Table 6.1 with an arbitrary sampling design.

Table 6.3: Point estimator and its variances of a person-level GREG estimator under simple cluster sampling

	Person-level GREG estimator
Assisting model ξ	$y_i = \boldsymbol{x_i}^T\boldsymbol{\beta_p} + \epsilon_i$
Point estimator	$\hat{T}_{y_p}^{\text{GREG}} = \hat{T}_{y_p}^{\text{HT}} + \boldsymbol{B_p}^T(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}})$
Variance	$V(\hat{T}_{y_p}^{\text{GREG}}) = \frac{M^2}{m}\left(1-\frac{m}{M}\right)\frac{1}{M-1}\sum_{g\in U_h}\left(\sum_{i\in U_g} r_i^{B_p}\right)^2$ with $r_g^{B_p} = y_g - \boldsymbol{x_g}^T\boldsymbol{B_p}$
Minimization problem	$\min_{\boldsymbol{B_p}}\sum_{g\in U_h}\sum_{i\in U_g}\left(r_i^{B_p}\right)^2$ with $r_i^{B_p} = y_i - \boldsymbol{x_i}^T\boldsymbol{B_p}$
Resulting coefficient	$\boldsymbol{B_p} = (\sum_{g\in U_h}\sum_{i\in U_g}\boldsymbol{x_i}\boldsymbol{x_i}^T)^{-1}\sum_{g\in U_h}\sum_{i\in U_g}\boldsymbol{x_i}y_i$

Table 6.3 clarifies the reason for the mismatch between the residuals in the minimization problem (fourth row) and the variance formula (third row). It is given by the order of the sum and the power of two, which can be seen from

$$\sum_{g\in U_h}\sum_{i\in U_g}(r_i^{B_p})^2 \neq \sum_{g\in U_h}\left(\sum_{i\in U_g} r_i^{B_p}\right)^2. \tag{6.1}$$

The left-hand side in (6.1) describes the minimization problem, the right-hand side the variance formula. We expect that the mismatch in (6.1) increases with the household size because it is valid that $(\sum_{i\in U_g} r_i^{B_p})^2 > \sum_{i\in U_g} {r_i^{B_p}}^2$. This inequality implies that the mismatch in (6.1) depends on U_g and thereby on N_g, the number of persons in household g.

To conclude, under cluster sampling the residuals in the minimization problem, which delivers the point estimator, contradict the residuals used in the corresponding variance formula. In other words, the variance applied to person-level GREG estimators does not conform to its underlying person-level model. In a simulation study, we examine whether the mismatch affects the precision of the variance estimates and whether it depends on the household sizes.

6.1.2 Optimal Estimator at the Person Level

Following Montanari (1987), the optimal estimator has minimum variance in a large class of estimators (see Section 2.3.1 for details on the optimal theory). Setting the first derivative of the variance of the person-level GREG estimator (third row in Table 6.2) to zero,

$$\begin{aligned} &\frac{\partial V(\hat{T}_{y_p}^{\text{GREG}})}{\partial \boldsymbol{B_p}} \overset{!}{=} 0 \\ \Leftrightarrow\ &\boldsymbol{B_p}^{\text{OPT}} = \Big(\sum_{g\in U_h} \boldsymbol{x_g}{\boldsymbol{x_g}}^T\Big)^{-1} \sum_{g\in U_h} \boldsymbol{x_g} y_g, \end{aligned} \tag{6.2}$$

yields the optimal coefficient. Accordingly, the optimal coefficient (6.2) depends on $\boldsymbol{x_g}$ and y_g instead on $\boldsymbol{x_i}$ and y_i as used in the assisting model ξ (first row in Table 6.2). OPT indicates optimal estimator. Hence, at the person level the optimal GREG estimator is given by

$$\hat{T}_{y_p}^{\text{OPT}} = \hat{T}_{y_p}^{\text{HT}} + {\boldsymbol{B_p}^{\text{OPT}}}^T(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}). \tag{6.3}$$

As a result, from the optimal point of view, one should always aggregate the available person-level variables and only use its household totals, even if the variable of interest is a person characteristic. However, the general choice of (6.3) is critical, in particular when the households tend to be heterogeneous. This problem is expected to be aggravate for clusters larger than households, such as in area cluster sampling. If the areas also are diversified, for example, by a mixture of social buildings and detached houses, and the variable of interest is volatile, such as income, the total estimates of income might be inaccurate.

A further problem of the general choice of (6.3) arises from the fact that following Robinson (1950), the correlations for the same variable computed at the person or at the household level can differ. Hence, aggregating the person-level information per household can lead to wrong conclusions about the true relationship between the auxiliaries and the variable of interest, resulting in an incorrect coefficient in the optimal GREG estimator. In the literature, the wrong inference is known as ecological fallacy (see Section 3.2.2 for details). The incorrect coefficient affects the efficiency of the optimal estimator since even if the GREG estimator is asymptotically unbiased, regardless of the correctness of the assisting model, its efficiency depends on the explanatory power of the model. Thus, in cases of ecological fallacy, we doubt the superiority of the optimal estimator compared with a person-level GREG estimator.

Moreover, the auxiliaries $\boldsymbol{x_g}$ of the optimal estimator (6.3) do not contain an intercept. Consequently, the resulting weights do not sum up to the number of persons in the population.

To conclude, we expect that the optimality of (6.2) is influenced by the size and heterogeneity of the households and by the correlation between the auxiliaries and the variable of interest at both levels. In the following simulation study, we verify these expectations. For a comparison, table 6.3 summarizes the person-level GREG estimator against the optimal GREG estimator under cluster sampling.

Table 6.4: Person-level and optimal GREG estimator under cluster sampling

Person-level GREG estimator	Optimal GREG estimator
$\hat{T}_{y_p}^{\text{GREG}} = \hat{T}_{y_p}^{\text{HT}} + \boldsymbol{B_p}^T(\boldsymbol{T_x} - \boldsymbol{\hat{T}_x^{\text{HT}}})$	$\hat{T}_{y_p}^{\text{OPT}} = \hat{T}_{y_p}^{\text{HT}} + \boldsymbol{B_p^{\text{OPT}}}^T(\boldsymbol{T_x} - \boldsymbol{\hat{T}_x^{\text{HT}}})$
$\boldsymbol{B_p} = (\sum_{g \in U_h} \sum_{i \in U_g} \boldsymbol{x_i}\boldsymbol{x_i}^T)^{-1} \sum_{g \in U_h} \sum_{i \in U_g} \boldsymbol{x_i} y_i$	$\boldsymbol{B_p^{\text{OPT}}} = \Big(\sum_{g \in U_h} \boldsymbol{x_g}\boldsymbol{x_g}^T\Big)^{-1} \sum_{g \in U_h} \boldsymbol{x_g} y_g$
$V(\hat{T}_{y_p}^{\text{GREG}}) = \sum_{g \in U_h} \sum_{k \in U_h} \triangle_{gk} \frac{r_g^{B_p}}{\pi_g} \frac{r_k^{B_p}}{\pi_k}$ with $r_g^{B_p} = y_g - \boldsymbol{x_g}^T \boldsymbol{B_p}$	$V(\hat{T}_{y_p}^{\text{OPT}}) = \sum_{g \in U_h} \sum_{k \in U_h} \triangle_{gk} \frac{r_g^{B_p^{\text{OPT}}}}{\pi_g} \frac{r_k^{B_p^{\text{OPT}}}}{\pi_k}$ with $r_g^{B_p^{\text{OPT}}} = y_g - \boldsymbol{x_g}^T \boldsymbol{B_p^{\text{OPT}}}$

6.2 Literature on Alternative Variance Formulas for GREG Estimators under Cluster Sampling

The previously discussed consequences of the variance formula under cluster sampling for person-level GREG estimators strongly depend on the aggregated form of the variance formula. Therefore, in this section, we briefly review the literature on alternative variance formulas to the design-based variance formula. In the model-based context, Royall (1992) examined the best linear unbiased (BLU) estimator and derived its model-variance. As already outlined in Section 2.2.1, in the model-based approach point and variance estimators are motivated by a working model. When the working model is assumed to be linear, the BLU estimator is equivalent to the GREG estimator. By relating the BLU estimator to the Horvitz-Thompson estimator Royall (1992), built a bridge to the design-based approach. Tam (1995) extended the BLU estimator to cluster sampling and assumed that the covariance matrix of the working model is block-diagonal. However, since the block-diagonal covariance structure results in the same per-household aggregation of person-level information as the design-based variance formula (cf. Särndal et al., 1992, p. 307), we do not pursue the BLU estimator and its model-variance in the following. Therefore, the critical disregard of the initial level of modeling remains unchanged.

Valliant (2002) offered a leverage-adjusted sandwich estimator to estimate the model variance of a GREG estimator. The sandwich variance estimator consists of the squared residuals that are adjusted by factors analogous to leverages known from econometrics. He showed that the proposed sandwich estimator is approximately model- and design-unbiased. Kennel (2013) extended the leverage-adjusted sandwich variance estimator to cluster sampling designs. However, the sandwich variance estimator is also based on household total residuals. Therefore, to the best of our knowledge, there are no alternative variance formulas discussed in the literature, which prevent the per-household aggregation of person-level information.

6.3 Proposed Hybrid GREG Estimator

As declared in Section 6.1, we assess the general choice of the optimal estimator (6.3), which utilizes the per-household aggregates of the variables to estimate person-level characteristics as critical, especially for large and heterogeneous households. As a remedy, we develop a hybrid GREG estimator that compromises between the optimal and the person-level GREG estimator. The proposed hybrid GREG estimator is implemented at the person level, since the variables of interest are person-level characteristics. This proceeding inhibits ecological fallacy. The intention of the hybrid GREG estimator is to incorporate the auxiliary information of the household members additional to the information of the persons. We define the hybrid GREG estimator as

$$\hat{T}_{y_p}^{\text{HYB}} = \hat{T}_{y_p}^{\text{HT}} + \boldsymbol{B}_{\boldsymbol{p}}^{\text{HYB}^T}(\boldsymbol{T_x} - \hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}) \tag{6.4}$$

where the coefficient is given by

$$\boldsymbol{B}_{\boldsymbol{p}}^{\text{HYB}} = \left(\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g}\tilde{\alpha}_{ij}\boldsymbol{x_i}\boldsymbol{x_j}^T\right)^{-1}\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g}\tilde{\alpha}_{ij}\boldsymbol{x_i}y_j \tag{6.5}$$

with weighting factors

$$\tilde{\alpha}_{ij} = \begin{cases} \alpha_g & \text{for } i = j \\ (1-\alpha_g) & \text{for } i \neq j. \end{cases}$$

Abbreviation HYB indicates hybrid GREG estimator. The double sum in the coefficient (6.5) allows us to utilize both the person- and household-level information of the auxiliaries and the variable of interest. For explanation of the double sum and the weighting factor, consider the case of $Q = 2$ auxiliary variables. Then, the first term in (6.5) is determined by

$$\left(\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g}\tilde{\alpha}_{ij}\boldsymbol{x_i}\boldsymbol{x_j}^T\right)^{-1} = \begin{pmatrix} \sum\limits_{g\in U_h}\sum\limits_{i\in U_g}\sum\limits_{j\in U_g}\tilde{\alpha}_{ij}x_{i1}x_{j1} & \sum\limits_{g\in U_h}\sum\limits_{i\in U_g}\sum\limits_{j\in U_g}\tilde{\alpha}_{ij}x_{i2}x_{j1} \\ \sum\limits_{g\in U_h}\sum\limits_{i\in U_g}\sum\limits_{j\in U_g}\tilde{\alpha}_{ij}x_{i1}x_{j2} & \sum\limits_{g\in U_h}\sum\limits_{i\in U_g}\sum\limits_{j\in U_g}\tilde{\alpha}_{ij}x_{i2}x_{j2} \end{pmatrix}^{-1} \tag{6.6}$$

with first diagonal element

$$\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g}\tilde{\alpha}_{ij}x_{i1}x_{j1}=\sum_{g\in U_h}\alpha_g\underbrace{\sum_{i\in U_g}x_{i1}^2}_{\text{Info of the persons}}+\sum_{g\in U_h}(1-\alpha_g)\underbrace{\sum_{i\in U_g}\sum_{\substack{j\in U_g\\ j\neq i}}x_{i1}x_{j1}}_{\text{Info with all other household members}}$$

and with minor diagonal elements

$$\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g}\tilde{\alpha}_{ij}x_{i1}x_{j2}=\sum_{g\in U_h}\alpha_g\underbrace{\sum_{i\in U_g}x_{i1}x_{i2}}_{\text{Info of the persons}}+\sum_{g\in U_h}(1-\alpha_g)\underbrace{\sum_{i\in U_g}\sum_{\substack{j\in U_g\\ j\neq i}}x_{i1}x_{j2}}_{\text{Info with all other household members}}.$$

Accordingly, the diagonal elements in (6.6) concerns the auxiliary information of both the persons and of all other household members with respect to the same variable. The minor diagonal elements contain the cross-auxiliary information of the persons and the other household members. The weighting factors α_g and $(1-\alpha_g)$ allow us to differently weight information of the persons and of other household members.

The second term in (6.5) can be rewritten as

$$\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g}\tilde{\alpha}_{ij}\boldsymbol{x_i}y_j=\begin{pmatrix}\sum\limits_{g\in U_h}\sum\limits_{i\in U_g}\sum\limits_{j\in U_g}\tilde{\alpha}_{ij}x_{i1}y_j\\ \sum\limits_{g\in U_h}\sum\limits_{i\in U_g}\sum\limits_{j\in U_g}\tilde{\alpha}_{ij}x_{i2}y_j\end{pmatrix}\tag{6.7}$$

with first row element

$$\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g}\tilde{\alpha}_{ij}x_{i1}y_j=\sum_{g\in U_h}\alpha_g\underbrace{\sum_{i\in U_g}x_{i1}y_i}_{\text{Info of the persons}}+\sum_{g\in U_h}(1-\alpha_g)\underbrace{\sum_{i\in U_g}\sum_{\substack{j\in U_g\\ j\neq i}}x_{i1}y_j}_{\text{Info with all other household members}}$$

and second row element

$$\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g}\tilde{\alpha}_{ij}x_{i2}y_j=\sum_{g\in U_h}\alpha_g\underbrace{\sum_{i\in U_g}x_{i2}y_i}_{\text{Info of the persons}}+\sum_{g\in U_h}(1-\alpha_g)\underbrace{\sum_{i\in U_g}\sum_{\substack{j\in U_g\\ j\neq i}}x_{i2}y_j}_{\text{Info with all other household members}}.$$

To make the difference between the hybrid GREG estimator (6.3) and a person-level GREG estimator,

$$\hat{T}_{y_p}=\hat{T}_{y_p}^{\text{HT}}+\boldsymbol{B_p}^T(\boldsymbol{T_x}-\hat{\boldsymbol{T}}_{\boldsymbol{x}}^{\text{HT}}),\tag{6.8}$$

more obvious, we express the coefficient

$$\boldsymbol{B_p}=\Big(\sum_{g\in U_h}\sum_{i\in U_g}\boldsymbol{x_i}\boldsymbol{x_i}^T\Big)^{-1}\sum_{g\in U_h}\sum_{i\in U_g}\boldsymbol{x_i}y_i$$

in a form comparable to that of the hybrid GREG estimator. For $Q = 2$, the coefficient $\boldsymbol{B_p}$ is obtained from

$$\left(\sum_{g \in U_h} \sum_{i \in U_g} \boldsymbol{x_i}\boldsymbol{x_i}^T\right)^{-1} = \begin{pmatrix} \sum\limits_{g \in U_h} \sum\limits_{i \in U_g} x_{i1}^2 & \sum\limits_{g \in U_h} \sum\limits_{i \in U_g} x_{i2}x_{i1} \\ \sum\limits_{g \in U_h} \sum\limits_{i \in U_g} x_{i1}x_{i2} & \sum\limits_{g \in U_h} \sum\limits_{i \in U_g} x_{i2}^2 \end{pmatrix}^{-1}$$

and

$$\sum_{g \in U_h} \sum_{i \in U_g} \boldsymbol{x_i} y_i = \begin{pmatrix} \sum\limits_{g \in U_h} \sum\limits_{i \in U_g} x_{i1}y_i \\ \sum\limits_{g \in U_h} \sum\limits_{i \in U_g} x_{i2}y_i \end{pmatrix}.$$

Hence, the person-level GREG estimator (6.8) captures only the information of the persons itself, but not the cross-household information. The main differences between $\boldsymbol{B_p^{\text{HYB}}}$ and $\boldsymbol{B_p}$ are driven by the second terms in (6.6) and (6.7) and by the weighting factor $\tilde{\alpha}_{ij}$.

Different choices of weighting factors are possible. One possible choice is that the weighting factors are chosen to account for the heterogeneity of the households because as explained in Section 6.1.2 the optimality of the GREG estimator depends on it. A measure of the heterogeneity is the within variance, given by

$$V^{\text{within}}(y) = (M-1) \sum_{g \in U_h} \sum_{i \in U_g} (y_i - \bar{y}_i)^2$$

with $\bar{y}_i = N^{-1} \sum_{i \in U_p} y_i$ as mean value. Then, we define

$$\alpha_g = \begin{cases} 1 & \text{for } \ g \in \{1, \ldots, M : N_g = 1\} \\ \dfrac{\sum_{i \in U_g} (y_i - \bar{y}_i)^2}{\sum_{g \in U_h} \sum_{i \in U_g} (y_i - \bar{y}_i)^2} & \text{otherwise.} \end{cases} \tag{6.9}$$

Accordingly, α_g reflects the share of the within variance of household g on the total within variance $V^{\text{within}}(y)$. The more heterogeneous household g, the higher α_g that weights the information of the persons. For single-person households, there is no information from other household members; thus per definition, the weighting factor equals 1. It should be noted that a_g can becomes very small if the number of households is large.

Another possible choice of α_g is given by

$$\alpha_g = \begin{cases} 1 & \text{for } \ g \in \{1, \ldots, M : N_g = 1\} \\ (1 - \frac{1}{N_g}) & \text{otherwise.} \end{cases} \tag{6.10}$$

Therefore, (6.10) respects the household size as a further factor influencing the optimality of the GREG estimator. The intention behind of this choice is that the larger the household g, the lower $(1 - \alpha_g)$ that weights the information of all other household members. The advantage of (6.10) compared to (6.9) is that it is independent from the variable of interest.

The compromise between the optimality and person-level modeling of our proposed hybrid GREG estimator becomes clear for certain weighting factors. Thus, we show that for certain choices of the weighting factors either the optimal or the person-level GREG estimator results. Given the following equalities

$$\sum_{g\in U_h} \boldsymbol{x_g}\boldsymbol{x_g}^T = \sum_{g\in U_h} \Big(\sum_{i\in U_g} \boldsymbol{x_i}\Big)\Big(\sum_{i\in U_g} \boldsymbol{x_i}\Big)^T = \sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g} \boldsymbol{x_i}\boldsymbol{x_j}^T$$
$$\sum_{g\in U_h} \boldsymbol{x_g} y_g = \sum_{g\in U_h} \Big(\sum_{i\in U_g} \boldsymbol{x_i}\Big)\Big(\sum_{i\in U_g} y_i\Big) = \sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g} \boldsymbol{x_i} y_j,$$

it can be shown that for $\tilde{\alpha}_{ij} = 1$, our proposed coefficient (6.5) can be expressed as the optimal coefficient (6.2)

$$\begin{aligned} \boldsymbol{B_p}^{\text{HYB}} &= \left(\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g} \boldsymbol{x_i}\boldsymbol{x_j}^T\right)^{-1} \sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g} \boldsymbol{x_i} y_j \\ &= \Big(\sum_{g\in U_h} \boldsymbol{x_g}\boldsymbol{x_g}^T\Big)^{-1} \sum_{g\in U_h} \boldsymbol{x_g} y_g \\ &= \boldsymbol{B_p}^{\text{OPT}}. \end{aligned}$$

Consequently, for $\tilde{\alpha}_{ij} = 1$, the proposed hybrid GREG estimator simplifies to the optimal estimator

$$\hat{T}_{y_p}^{\text{HYB}} = \hat{T}_{y_p}^{\text{OPT}}.$$

On the other hand, for $\alpha_g = 1$ the information of the other household members in the double sums in (6.6) and (6.7) is weighted by zero. In result, the hybrid GREG estimator equals the person-level GREG estimator

$$\hat{T}_{y_p}^{\text{HYB}} = \hat{T}_{y_p}^{\text{GREG}}.$$

Therefore, the optimal and the person-level GREG estimator can be seen as special cases of the hybrid GREG estimator.

An estimator of the coefficient (6.5) is given by

$$\hat{\boldsymbol{B}}_{\boldsymbol{p}}^{\text{HYB}} = \left(\sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g} \frac{\tilde{\alpha}_{ij}\boldsymbol{x_i}\boldsymbol{x_j}^T}{\pi_i}\right)^{-1} \sum_{g\in U_h}\sum_{i\in U_g}\sum_{j\in U_g} \frac{\tilde{\alpha}_{ij}\boldsymbol{x_i} y_j}{\pi_i}.$$

The variance of the proposed hybrid GREG estimator (6.4) is estimated by the residual variance

$$V(\hat{T}_{y_p}^{\text{HYB}}) = \sum_{g\in U_h}\sum_{k\in U_h} \triangle_{gk} \frac{r_g^{B_p^{\text{HYB}}}}{\pi_g} \frac{r_k^{B_p^{\text{HYB}}}}{\pi_k}$$

with $r_g^{B_p^{\text{HYB}}} = \sum_{i \in U_g} (y_i - \boldsymbol{x_i}^T \boldsymbol{B}_{\boldsymbol{p}}^{\text{HYB}})$ as residual.

In conclusion, our proposed hybrid GREG estimator is derived at the person level but additionally includes the information of the other household members. The receptive extent to which the information of the persons and the households is incorporated is determined by the weighting factors. The nearer a_g is to 1, the more similar our proposed hybrid GREG estimator to the person-level GREG estimator. The nearer α_{ij} is to 1, the more similar our proposed hybrid GREG estimator to the optimal estimator.

6.4 Simulation Study

The simulation study is based on the same simulation setup as introduced in Section 3.4.1. The objective of the simulation study is twofold: First, in Section 6.4.1, we examine the consequences of the aggregated form of the variance formula under cluster sampling for person-level GREG estimators. Second, in Section 6.4.2, we compare point and precision estimates of our proposed hybrid GREG estimator with the optimal and the person-level GREG estimator. The estimators under consideration are presented in Table 6.5.

Table 6.5: Estimators under consideration

Estimator	Description
PERS	Person-level GREG estimator defined in (6.8)
OPT	Optimal GREG estimator defined in (6.3)
HYBa	Hybrid GREG estimator (6.4) with weighting factors (6.9)
HYBb	Hybrid GREG estimator (6.4) with weighting factors (6.10)

As explained in Section 6.1.2, we expect that the performance of the estimators is influenced by the household size, by the heterogeneity of the households, and by the correlation between the auxiliaries and the variable of interest at both levels. To study the influence, we conduct different scenarios that are summarized in Table 6.6.

Table 6.6: Scenarios with different household decompositions

Scenario	Description
HETEROGENEOUS	Original household IDs from AMELIA
HOMOGENEOUS	Generated household IDs depending on income
HETEROGENEOUS_large	Collapsed original household IDs from AMELIA
HOMOGENEOUS_large	Collapsed generated household IDs depending on income

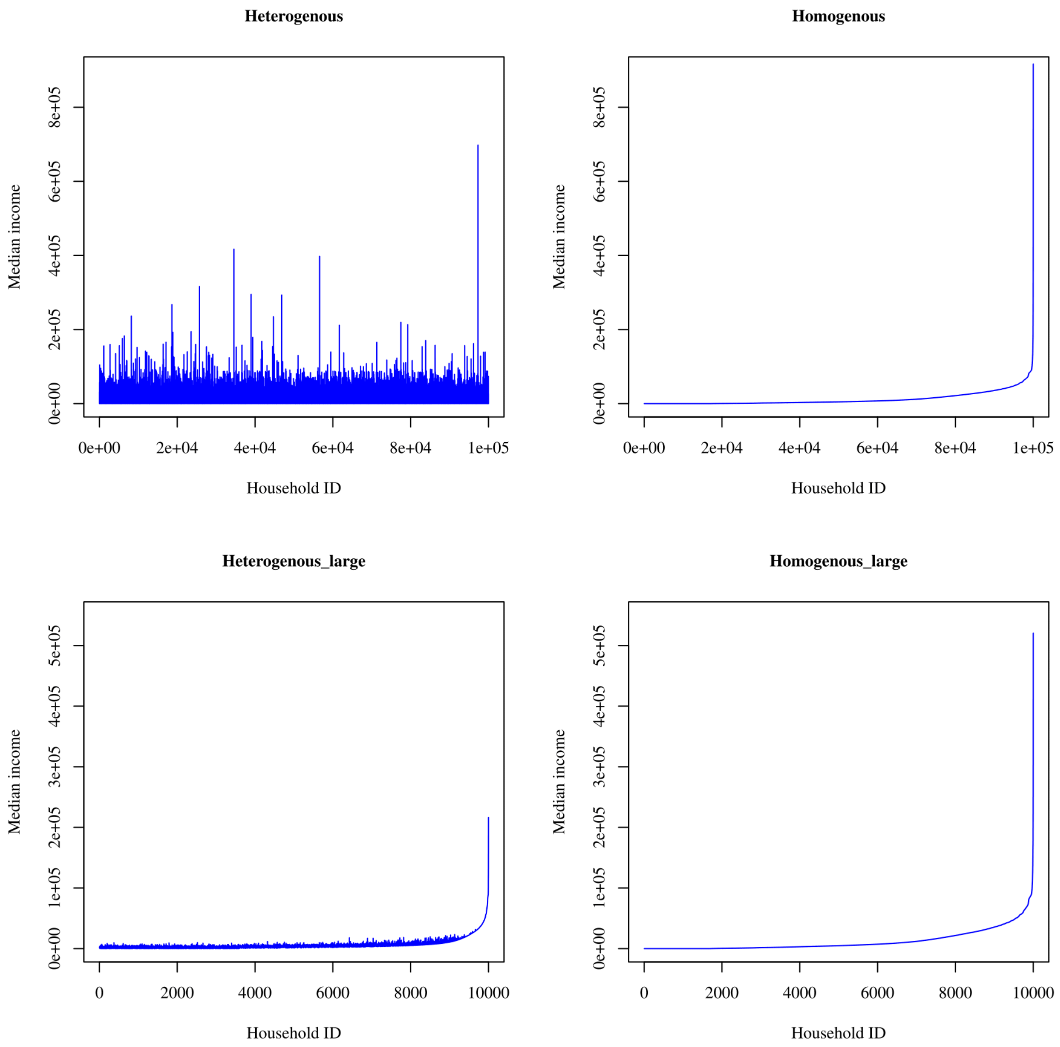

Figure 6.1: Boxplots of the median income of the households under different scenarios

In order to study the influence of the heterogeneity of the households, we generate households of different decompositions. The base case is given by the original household identifier (ID) from the AMELIA data set. Figure 6.1 illustrates the median income for every household under different scenarios. We plot the median, since it is robust against outliers. The upper left plot shows the base case. It can be seen that the median income of the original household IDs is very volatile. Consequently, the households are heterogeneous with respect to income. We denote this scenario as HETEROGENEOUS. To generate households that are very similar, on the other side, we redistribute the persons in our data set to new households. For this purpose, we sort the persons by income. Then, we generate a new household ID by randomly allocating the known distribution of the household sizes from the original household ID to the sorted persons. We call this scenario HOMOGENEOUS. As a result, the distributions of the household sizes are the same for HETEROGENEOUS and HOMOGENEOUS. The right plot in Figure 6.1 shows that the median income is similar for most households, except for the households with the highest income.

In order to examine whether the performance of the estimators is influenced by the household sizes, we increase the household size by collapsing 10 households of the original household ID to one larger cluster. We denote this scenario as HETEROGENEOUS_large. Analogously to the proceeding in the second scenario, we sort the persons by income and generate a new household ID by randomly allocating the known distribution of the household sizes from the collapsed original household IDs to the sorted persons. We call this scenario HOMOGENEOUS_large. The resulting household size distributions of HETEROGENEOUS_large and HOMOGENEOUS_large are the same. The median income of the households is illustrated in the lower plots in Figure 6.1. Of course, the total number of the households in the lower plots is decreased by a factor 10 compared with the upper plots.

Finally, to explore the effect of the divergent strength of the relationship between the auxiliaries and the variable of interest computed at the person or household level, we choose `sex` as the auxiliary variable. As mentioned in Section A.2 in Appendix A, the sign of the correlation between `sex` and `inc`, computed at the person or household level, differs. The signs of the correlations between `sex` and the other variables of interest are the same at both levels.

We draw $R = 1000$ MC samples of $m = 200$ households in the first two scenarios and $m = 20$ households in the last two scenarios. Thus, the resulting sample sizes of persons is approximately the same in all four scenarios. To reduce the number of plots in the figures, we select `inc`, `soc`, `sel`, `act_2` and `bene_age2` from Table 3.7 as variables of interest. The results for the remaining variables of interest are very similar.

The coefficients, RB and rsRB$_r$ of point and variance estimates obtained from the estimators under consideration are presented in Figures 6.2, 6.3 and 6.4. The mean values are indicated in green. All figures confirm that the coefficients and the point and variance estimates vary less in the heterogeneous than in the homogeneous scenarios. This is the expected result under cluster sampling, but nevertheless, our main concerns are addressed in the following two sections.

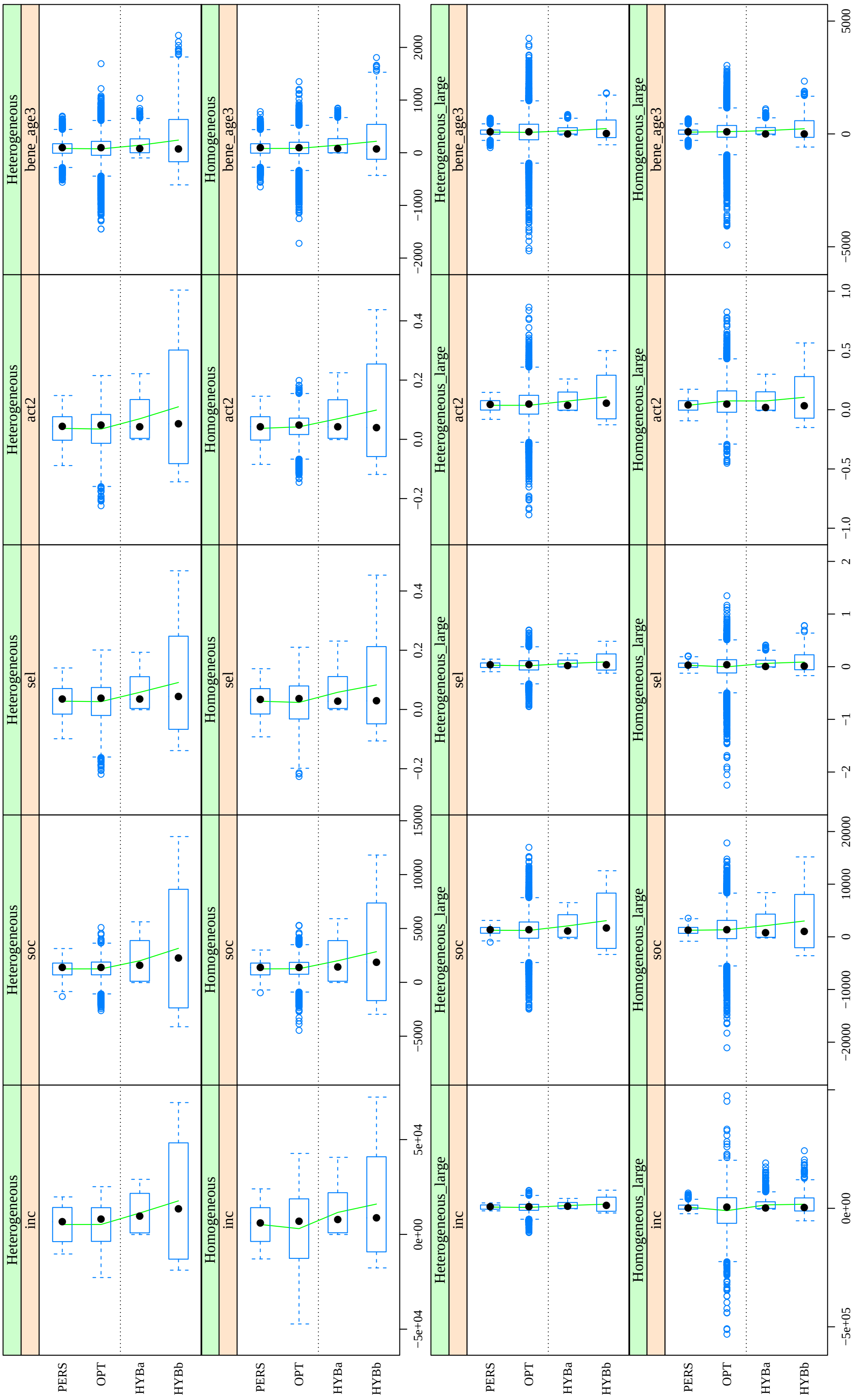

Figure 6.2: Coefficients under different scenarios

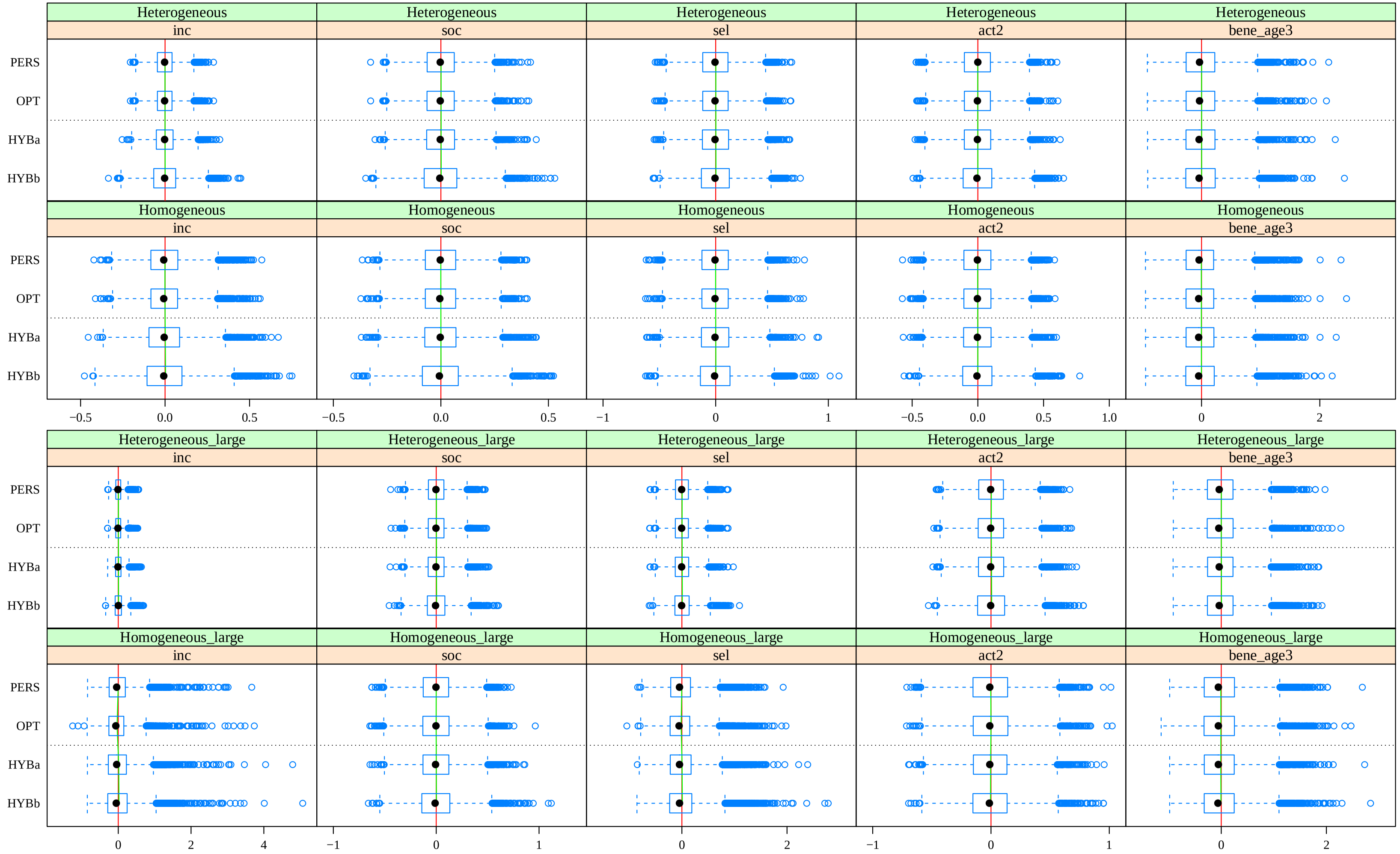

Figure 6.3: Relative bias and replicate specific-relative bias of the point estimates

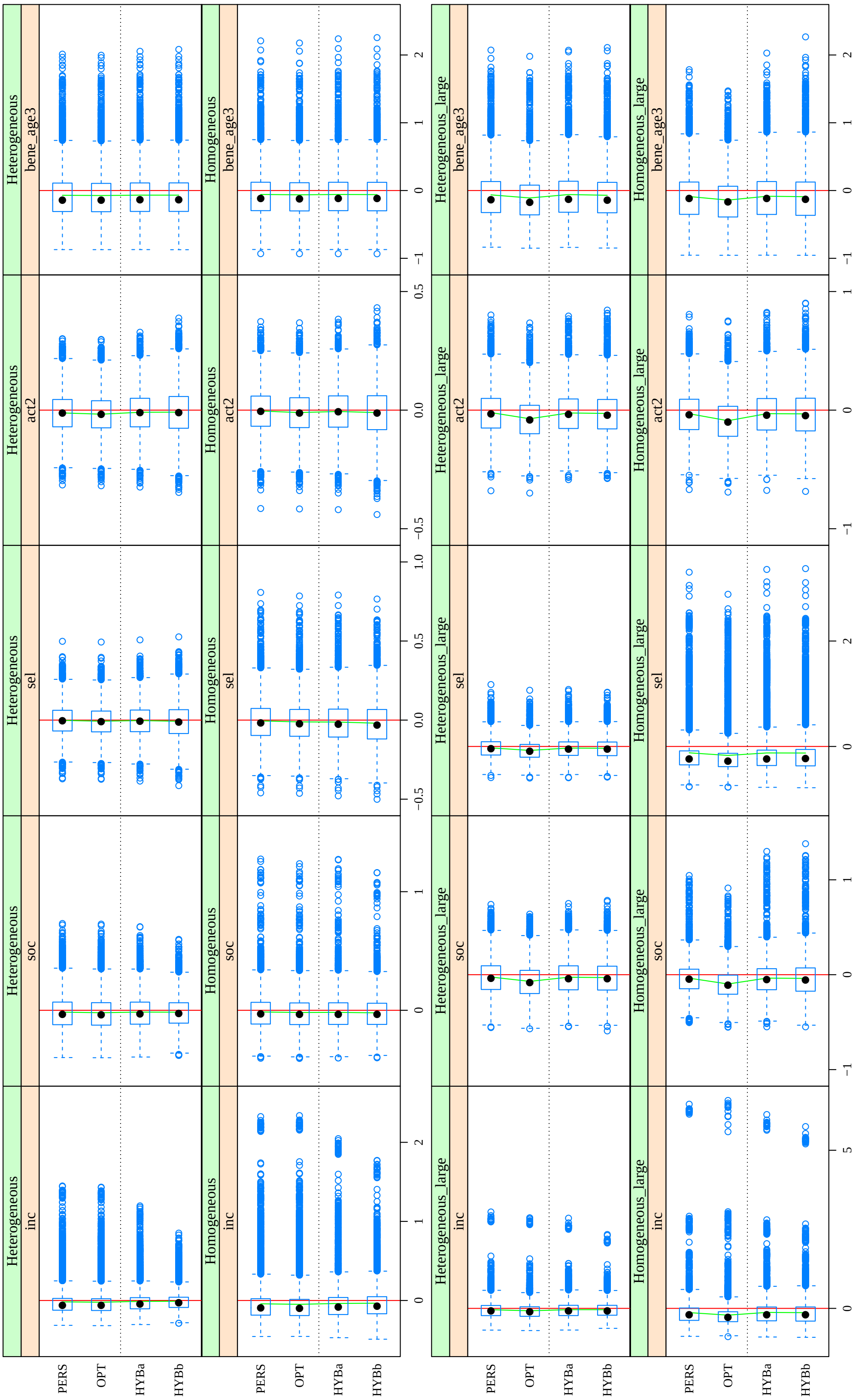

Figure 6.4: Relative bias and replicate specific-relative bias of the estimated variances

6.4.1 Consequences of the Variance on Person-Level GREG Estimators

For the discussion of the consequences of the aggregated form of the variance formula on person-level GREG estimators, we focus on the person-level GREG estimator PERS and the optimal GREG estimator OPT. In this section, we answer two questions:

First, are the variance estimates of the person-level GREG estimator affected by the mismatch between the residuals in the minimization problem and in the variance formula? To answer this question, we examine the RB and rsRB$_r$ for $r = 1, \ldots, 1000$ of the variance estimates in Figure 6.4 for PERS. The RB is highlighted in green. It can be seen that PERS tends to underestimate the empirical variance of the person-level GREG estimator, particularly for HOMOGENEOUS_large. This result is an indication that our expectation that the mismatch of the residuals affects the precision of the variance estimates for larger and homogeneous household, is confirmed. It should be noted that the rsRB$_r$ of `inc` is very large for some MC replicates. The reasons include that `inc` has a very skew distribution, HOMOGENEOUS and HOMOGENEOUS_large are the worst case scenarios under cluster sampling and for HOMOGENEOUS_large the sample size is very small ($m = 20$).

Second, is the optimality of OPT influenced by the heterogeneity of the households, the household size or the correlation between the variable of interest and the auxiliaries? Table 6.2 depicts that the ranges of the coefficients of OPT are considerably wider than the ranges of PERS, in particular for the scenarios with the larger household sizes in the lower plots. The wider range is caused by the differing signs of the correlations computed at the person and household level. With respect to the point estimates in Figure, 6.3 OPT and PERS perform similar. However, for HOMOGENEOUS_large, OPT slightly underestimates the true population total for `inc` and `sel`. Figure 6.4 shows that the variance estimates of OPT considerably underestimates the variance for both scenarios of larger household sizes (lower plots). These results indicate that the influence of the above factors on the point estimates of OPT compared to PERS is limited to the worst case scenario of large and homogeneous households. The point estimates also demonstrate that OPT is not necessarily superior to PERS, even if the former is the optimal estimator. Also with respect to the variance estimates OPT suffers from the larger household sizes and the heterogeneity. Therefore, we conclude that the decision between a person-level or an optimal GREG estimator should be considered with caution and should take into account the size and heterogeneity of the households.

6.4.2 Performance of the Hybrid GREG Estimator

In this section, we compare our proposed hybrid estimator with a person-level GREG estimator and an optimal GREG estimator. The hybrid estimator is implemented with two different weighting factors. HYBa is based on weighting factor (6.9) and accounts for the heterogeneity of the households. HYBb relies on weighting factor (6.10) and respects the household size. Figure 6.2 depicts that the coefficients of HYBb varies considerably more than the coefficients of HYBa. Since the coefficients are the only difference between the point estimators, we expect the same pattern of both estimators for the point estimates. Compared with OPT, the boxes of HYBa are smaller. The reverse is true compared with PERS. This observation is in accordance with the fact that our proposed estimator is a hybrid between the optimal and the person-level GREG estimator. For HYBb, this observation applies only for the scenarios with the larger household sizes.

Figure 6.3 confirms that the point estimates of HYBa and HYBb are unbiased for all variables and all scenarios. HYBa performs very similar compared with PERS and OPT. The boxes of HYBb are slighly wider. The differences between HYBa and HYBb are less than expected from the pattern in the coefficients. Table 6.7 presents the RRMSE of the point estimates. There is a tendency that the RRMSE of HYBa and HYBb exceeds the RRMSE of PERS and OPT. This results shows that further research on the optimization of the weighting factors is needed to improve the efficiency of the point estimates.

With respect to the variance estimates in Figure 6.4 both proposed hybrid GREG estimators, HYBa and HYBb, achieve the most precise results, in particular for `inc` and `soc`. Surprisingly, this superiority is stronger for smaller households.

To conclude, the choice of a weighting factor accounting for the heterogeneity (HYBa) seems to be preferable compared to a choice accounting for the household size (HYBb). This implies that the heterogeneity of the households or clusters is more relevant for the quality of point and variance estimates. With regard to point estimation, there is further need for improvement of to weighting factors, and with regard to variance estimation, our hybrid estimator is superior compared with the optimal GREG estimator and the person-level GREG estimator.

6.5 Summary and Conclusion

In this chapter, we explored the consequences of the per-household aggregation of the person-level variables in the variance formula under cluster sampling for person-level GREG estimators. A first consequence is that the residuals in the variance formula and the residuals that determine the point estimator differ. Our simulation study showed that the variances estimates for a person-level GREG estimator indeed tends to underestimate the empirical variance for cases of large and homogeneous households.

A second consequence is that the form of the variance formula leads to an optimal estimator that depends on the aggregates of both the variable of interest and the auxiliary variables, even if the objective to estimate is a person-level characteristic. Our results have shown that the optimal GREG estimator is not superior to a person-level GREG estimator. Furthermore, we found that for larger homogeneous households it achieves less precise point and variance estimates than a person-level GREG estimator. Therefore, we recommend to be careful with the general choice of an optimal GREG estimator, especially for large and heterogeneous households.

To compromise between the optimal and the person-level GREG estimator, we propose the hybrid GREG estimator that incorporates both the person- and household-level information. The

Table 6.7: Relative bias and relative root mean squared error of point estimates for different scenarios

	RRMSE			
	PERS	OPT	HYBa	HYBb
	Heterogeneous			
`inc`	0.06	0.06	0.07	0.10
`soc`	0.09	0.09	0.10	0.11
`sel`	0.17	0.17	0.17	0.18
`act2`	0.15	0.15	0.15	0.16
`bene_age3`	0.36	0.36	0.36	0.37
	Homogeneous			
`inc`	0.12	0.12	0.14	0.16
`soc`	0.10	0.10	0.11	0.12
`sel`	0.18	0.18	0.19	0.20
`act2`	0.15	0.15	0.16	0.17
`bene_age3`	0.36	0.36	0.37	0.37
	Heterogeneous_large			
`inc`	0.10	0.10	0.11	0.13
`soc`	0.11	0.12	0.12	0.13
`sel`	0.18	0.19	0.19	0.20
`act2`	0.16	0.16	0.16	0.17
`bene_age3`	0.37	0.38	0.37	0.38
	Homogeneous_large			
`inc`	0.36	0.34	0.41	0.44
`soc`	0.18	0.19	0.19	0.21
`sel`	0.31	0.31	0.33	0.36
`act2`	0.22	0.22	0.21	0.22
`bene_age3`	0.43	0.44	0.43	0.44

weighting factors determine the extent to which the person- and household-level information is included in the estimation. Thus, the weighting factors balance the compromise between the optimal and the person-level GREG estimator. The proposed hybrid GREG estimator is a very flexible method because due to the weighting factors, it can be adjusted to the specific sample conditions. The special case of $\tilde{\alpha_{ij}} = 1$ delivers the optimal GREG estimator, $\alpha_g = 1$ in turn results in the person-level GREG estimator. Therefore, the hybrid GREG estimator builds a bridge between the separated approaches of person- or household-level modeling. The simulation study verifies that the choice of a weighting factor accounting for the heterogeneity is preferable compared to a weighting factor accounting for the household size. Future research should deal with optimal weighting factors to increase the precision of point estimates. Conceivable further choices are intraclass correlation coefficients. We discourage from using design effects to weight the person- and household-level information, since the design effects depend on the variance formula under cluster sampling with the disadvantages discussed in Section 6.1.

The hybrid GREG estimator can also be used to ensure consistency between person- and household-level estimates. For this purpose, the common variables have to be included as additional auxiliaries into the estimator as proposed in Chapter 4.

7 Conclusion and Outlook

In this thesis, we focused on topics on consistent estimation in household surveys. In Chapter 3, we investigated integrated weighting as current practice in official statistics to ensure consistent estimates at the person and the household level. In order to compare the proposed integrated weighting approaches introduced by Lemaître and Dufour (1987) and Nieuwenbroek (1993) we combine both to one general integrated GREG estimator. The integrated GREG estimator ensures consistency by replacing the original auxiliary information by constructed household mean values. The person weights, which are equal for all household members, are assigned one-to-one the corresponding households. Due to the one-to-one weight assignment between the person and the household level an additional auxiliary is required that ensures that the person weights sum up to the number of persons in the population and simultaneously the household weights sum up to the number of households in the population. This property, which we denote as integrative property, is to the best of our knowledge neglected in the literature. To deduce the consequences of the strict requirement of equal weights for all persons within the same household and the household itself, we opposed the integrated GREG estimator with a naïve GREG estimator. As a result, the integrated GREG estimator is characterized by an increased number of outcome values of the auxiliaries, neglects the heterogeneity of the household, and raises possible problems induced by ecological fallacy. The simulation study confirmed that these consequences result in more spread weights, more varied coefficients, and less efficient point and variance estimates for smaller sample sizes compared with a naïve GREG estimator.

As an alternative to integrated weighting, in Chapter 4, we proposed two weighting approaches that ensure consistent estimates and allow for differing weights within a household. Consistency is guaranteed by incorporating the variables that are common to both the person and household level data sets as additional auxiliaries. Our first weighting approach is easier to implement, since only the household-level estimator is influenced by the common variables. The person-level estimator, in turn, remains unaffected by the consistency requirements. In the second weighting approach, both estimators incorporate the common variables. This enables us to produce the best available estimate for the unknown totals of the common variables. Therefore, in survey practice the choice for one of the weighting approaches should balance the reduced implementation effort of the first weighting approach with the improved quality of the estimates for the unknown common variable totals and the totals of variables related to them. The advantages of the alternative weighting approaches compared to integrated weighting are manifold. Firstly, consistency is ensured more directly and only for the relevant variables, instead of indirectly by aggregating the individual information per household. Secondly, using the original auxiliary information allows for divergent weights for the persons within the same household. Therefore, the heterogeneity in a household, if existing, is captured, and individual patterns are retained. Thirdly, separated weighting models can be implemented at the person and at the household level, which ensures more flexibility in variable selection and prevent from problems induced by ecological fallacy. Finally, no additional auxiliary variable is required to enforce the integrated property. Our simulation study strongly supports the superiority of our alternative weighting approaches relative to integrated weighting with respect to point and variance estimates. In particular, the second proposed weighting approach yields the most precise estimation results. The precision gains depend on the strength of the relation between the common variables and the variables of interest. As a result, we contradict the widespread perception in the literature that equal weights are required to ensure consistent estimates.

Future research should address the effects of consistency requirements on nonresponse adjustment. In general, methods to prevent a nonresponse bias proceed at the person level. Hence, the adjusted person weights are no longer necessarily equal within a household. In order to still guarantee consistency, Eurostat (cf. European Commission, 2014, p. 40) recommends averaging the adjusted person weights within a household and assign this average weight to all household members. In contrast, our alternative weighting approaches allow a nonresponse adjustment at the person level without the need for a subsequent averaging process of the resulting weights. The incorporation of the common variables guarantees consistency even in the case of nonresponse adjustment. Therefore, individual response patterns are retained. This flexibility reinforces the superiority of our alternative weighting approaches compared to integrated weighting.

The alternative weighting approaches are both expressed as GREG estimators. Since the GREG estimator is analytically representable, we are able to derive explicit formulas for point and variance estimators. The further advantage of the analytical expression is that we were able to decompose the point estimators into a naïve GREG estimator and an adjustment term to quantify the effects caused by the consistency requirements. However, all proposed GREG estimators can be embedded into the calibration estimation framework (as introduced in Section 2.3.4). An advantage of the calibration approach is that additional constraints can easily be implemented such as box constraints that ensure the weights are within certain bounds. Moreover, our alternative weighting approaches can be combined with the generalized calibration estimator proposed by Münnich et al. (2018). The intent of the generalized calibration estimator is to relax some constraints when the total number of constraints are very large or some constraints are measured with uncertainty for example by means of small area estimation methods. The combination of the alternative weighting approaches and the generalized calibration estimator is particularly useful when the number of variables required for consistency increases or consistency is required at different hierarchical levels.

Another application field for our proposed weighting approaches is integrated surveys such as the German Microcensus 2020. The aim of the German Microcensus 2020 is to integrate the household surveys in one survey with a common core sample and different subsamples (cf. Riede, 2013). Now, the alternative weighting approaches can be applied to produce consistent estimates between the core sample and the subsamples. Therefore, the alternative weighting approaches not only ensure consistency within one household surveys, they can further be extended to ensure consistency at a higher level in an integrated system. The only requirement is to determine the common variables for which consistency is desired.

In addition the practical applications for official statistics, this thesis contributes to the theoretical literature. In Chapter 5, we derived an efficiency comparison between a person-level GREG estimator and an integrated household-level GREG estimator. The difficulty was that both underlying assisting models are of different dimensions. The difference is constituted by the intercept in the household-level model. As a remedy, we decomposed the variance of an integrated GREG estimator into the variance of a reduced GREG estimator without an intercept, which is of the same dimension as a person-level GREG estimator, and an additional variance term that captures the effect of the intercept disregard by the reduced GREG estimator. For this purpose, we initially decomposed the integrated coefficients by applying mediation models known from psychology and sociology. Subsequently, we transferred the decomposition to the integrated residuals by the construction of pseudo-residuals that permit us to exactly quantify the effect of the intercept on the variance disregarded by the reduced household-level model. Finally, we extended our findings to the decomposition of the sum of squared residuals. This decomposition was inserted into the difference between the variances of a person-level GREG estimator and an integrated household-level GREG estimator. To assess the resulting difference, we deduced a relationship between the coefficients determining the differences. The idea was to exploit an overlap model that contains both auxiliaries determining the coefficients. As a result, the difference between both variances consists of two variance components. The first variance component is given by the variances of a person-level and a reduced household-level GREG estimator. It depends on the correlation between the original auxiliaries and the constructed household mean values. The second variance component considers the effect of the intercept disregarded by the reduced household-level model and is hardly predictable.

In Section 5.3, we applied our proposed decomposition to predict the difference between two coefficients of determination when adding or omitting explanatory variables. Our result permits a deeper understanding of how the implementation of additional explanatory variables causes supplementary explanatory power of the model. This application of our decomposition can be relevant for the variable selection process in econometrics and survey statistics. In survey statistics our decomposition delivers a criterion to decide which auxiliary variables (with known totals) should be included into the assisting model to increase the efficiency of the estimator.

In the last chapter, we explored the consequences of the variance formula under cluster sampling on the person-level GREG estimator. One consequence is that the form of the variance formula leads to an optimal estimator that depends on the aggregates of both the variable of interest and the auxiliary variables regardless of the level of the variable of interest. Therefore, there is a trade-off between person-level modeling as the appropriate level to estimate person-level characteristics and optimality induced by an estimator using the per-household aggregated variables. As a remedy, we introduced the hybrid GREG estimator that balances between a person-level and an optimal GREG estimator. The proposed hybrid GREG estimator is a very flexible method because due to the weighting factors, it can be adjusted to the specific sample conditions. Our simulation results showed that a weighting accounting for the heterogeneity achieves more precise point and variance estimates than a weighting factor accounting for the household sizes. Compared to a person-level and an optimal GREG estimator, our proposed hybrid GREG estimator is superior with respect to variance estimates. To improve the precision of the point estimates, further research to be conducted should include the optimal choice of the weighting factors to balance between person-level modeling and optimality. Conceivable choices of weighting factors are the intraclass correlation coefficients.

A Additional Material for Chapter 3

A.1 Additional Table for the Simulation Study

Table A.1: Relative bias of point estimates at the person level

	m=1500			m=200		
	GREG	INT1	INT2	GREG	INT1	INT2
inc	0.00	0.00	0.00	-0.00	0.00	-0.00
soc	0.00	0.00	0.00	0.00	0.00	0.00
sel	-0.00	-0.00	-0.00	0.00	0.00	0.00
act1	0.00	0.00	0.00	0.00	0.00	0.00
act2	0.00	0.00	0.00	0.00	0.00	0.00
act3	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
inc_hs1	-0.00	-0.01	-0.00	0.01	-0.01	-0.00
inc_hs2	-0.00	-0.00	-0.00	0.01	0.01	0.00
inc_hs3	0.00	0.00	0.00	-0.00	0.00	-0.00
inc_hs4	0.00	0.00	0.00	-0.00	-0.00	0.00
inc_hs5	-0.00	-0.00	0.00	-0.00	0.00	0.00
inc_hs6	-0.01	-0.01	-0.01	-0.00	-0.01	-0.01
bene_age1	0.01	0.01	0.01	0.01	0.01	0.01
bene_age2	0.01	0.01	0.01	0.02	0.02	0.02
bene_age3	0.00	0.00	0.00	0.02	0.01	0.01
bene_age4	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02

A.2 Further Simulation Results

While exploring the consequences of integrated weighting, we generate several further simulation results. We start by estimating certain subgroups and domains. Next, we analyze whether the integrated GREG estimator suffers from ecological fallacy. Moreover, we investigate whether the enforcement of equal weights for all household members leads to faulty inferences in regressions.

First, in practice, estimates for subgroups or domains are often of as much interest as population totals. Thus, we analyzed whether the estimates for specific subgroups suffer from the replacement of the original auxiliaries. As subgroups, we chose various cross-classifications of sex, age, ms, hs, and inc. Surprisingly, we observe neither a higher RB (Table A.2) nor a higher MSE A.3 compared to the results given in Tables A.1 and 3.8.

Table A.2: Relative bias of point estimates for domains

	m=1500			m=200		
	GREG	INT1	INT2	GREG	INT1	INT2
age4_sex1_ms2	0.00	0.00	0.00	0.00	0.00	0.00
age2_sex0_ms4	-0.00	-0.01	-0.00	-0.02	-0.02	-0.02
age4_sex0_ms2	-0.00	-0.00	0.00	0.01	0.00	0.01
age2_sex1_ms4	0.00	0.01	0.01	-0.02	-0.03	-0.02
age4_sex1_ms4	-0.00	-0.00	-0.00	0.00	0.00	0.00
act1_sex1	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
act1_hs1	0.00	0.00	0.00	0.01	-0.02	-0.01
act1_sex0	0.00	0.00	0.00	0.00	0.00	0.00
act1_hs6	-0.00	-0.00	-0.00	0.00	-0.00	-0.00
act2_hs1_age1	-0.01	-0.01	-0.01	-0.02	-0.02	-0.05
act1_hs1_age4	0.02	0.01	0.01	0.03	0.00	0.02
act1_hs6_age1	0.02	0.02	0.02	-0.01	-0.01	0.00
inc_ms1	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
inc_ms2	0.00	0.00	0.00	-0.00	0.00	0.00
inc_ms3	-0.00	-0.00	-0.00	0.01	0.01	0.01
inc_ms4	0.00	0.00	0.00	-0.00	-0.00	-0.00

Second, we observe that for inc and sex the sign of the true correlations actually differs between the person and household level

$$\mathrm{Cor}(\boldsymbol{x}_i, \boldsymbol{y}_i) = -0.10 \quad \text{versus} \quad \mathrm{Cor}(\boldsymbol{x}_g, \boldsymbol{y}_g) = 0.32.$$

Therefore, the tacit assumption in the integrated weighting approach that the relationship between the auxiliary variable, here sex, and the variable of interest, here inc, are the same, causes ecological fallacy as described by Robinson (1950) (see Section 3.2.2 for details). The question arises whether the allegedly *wrong* sign of the correlation at the household level, which

Table A.3: Relative efficiency of the mean squared error of point estimates for domains

	m=1500				m=200			
	$\frac{\text{INT1}}{\text{GREG}}$	$\frac{\text{INT2}}{\text{GREG}}$	$\frac{\text{INT1}}{\text{GREG2}}$	$\frac{\text{INT2}}{\text{GREG2}}$	$\frac{\text{INT1}}{\text{GREG}}$	$\frac{\text{INT2}}{\text{GREG}}$	$\frac{\text{INT1}}{\text{GREG2}}$	$\frac{\text{INT2}}{\text{GREG2}}$
age4_sex1_ms2	0.99	0.99	0.99	0.99	1.02	1.03	1.03	1.01
age2_sex0_ms4	1.01	1.01	1.01	1.00	1.03	1.04	1.03	1.03
age4_sex0_ms2	1.01	1.00	1.00	1.01	1.03	1.03	1.03	1.03
age2_sex1_ms4	1.00	0.99	0.99	1.00	1.01	1.01	1.01	1.00
age4_sex1_ms4	1.00	1.00	1.00	1.00	1.02	1.02	1.02	1.01
act1_sex1	1.00	1.00	1.00	1.00	1.02	1.03	1.03	1.02
act1_hs1	0.93	0.87	0.95	1.21	0.97	0.88	0.96	1.30
act1_sex0	1.00	1.00	1.00	1.00	1.02	1.03	1.03	1.00
act1_hs6	0.93	0.90	0.97	1.14	0.96	0.98	1.04	1.13
act2_hs1_age1	1.02	1.00	1.00	1.02	1.14	0.98	0.99	1.14
act1_hs1_age4	1.02	0.98	0.99	1.04	1.07	0.98	1.00	1.08
act1_hs6_age1	0.99	0.99	1.00	1.02	1.01	1.07	1.08	0.99
inc_ms1	1.00	1.00	1.00	1.00	1.02	1.02	1.02	1.01
inc_ms2	1.00	1.00	1.00	1.00	1.04	1.05	1.05	1.03
inc_ms3	1.01	1.00	1.00	1.00	1.03	1.02	1.01	1.01
inc_ms4	1.00	1.00	1.00	1.00	1.03	1.02	1.03	1.02

determines the sign of the coefficients within the estimators, introduces some bias. To answer this question, we calculate GREG, INT1, and INT2 for inc with sex as a single auxiliary variable. Furthermore, we compute INT1b and INT2b, which do not contain the integrated variable, N_g^{-1}, as an additional auxiliary. From Table A.4 in the Appendix A, it becomes apparent that even if the integrated GREG estimators uses the coefficient with the *wrong* signs, no biased estimates results. Moreover, the MSEs are equal for GREG, INT1, and INT2. However, when excluding the integrated variable from the auxiliaries, the point estimates of INT1b and INT2b are considerably less efficient than those produced by GREG. In other words, including the integrated variable as additional auxiliary reverses the negative effect of the wrong sign on the efficiency in cases of ecological fallacy.

Table A.4: Relative bias of point estimates with sex as auxiliary for different sample sizes

	m=1500					m=200				
	GREG	INT1	INT2	INT1b	INT2b	GREG	INT1	INT2	INT1b	INT2b
inc	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table A.5: Relative efficiency of the MSE of the point estimates with solely sex as auxiliary

	m=1500				m=200			
	$\frac{\text{INT1}}{\text{GREG}}$	$\frac{\text{INT2}}{\text{GREG}}$	$\frac{\text{INT1b}}{\text{GREG}}$	$\frac{\text{INT2b}}{\text{GREG}}$	$\frac{\text{INT1}}{\text{GREG}}$	$\frac{\text{INT2}}{\text{GREG}}$	$\frac{\text{INT1b}}{\text{GREG}}$	$\frac{\text{INT2b}}{\text{GREG}}$
inc	1.00	1.00	1.10	1.10	1.00	1.00	1.08	1.08

To further analyze whether the wrong sign of a coefficient, and thus ecological fallacy, might produce some biased results, we generated several synthetic populations with different correlation structures between y_i and x_i or $\bar{x}_i$. Our results show once more that even when Cor(y, x) and Cor$(y, \bar{x})$ have different signs, the integrated GREG estimator does not perform significantly worse than a naïve GREG estimator with respect to RB and MSE when the integrated variable is included. Therefore, we refrain from tabulating the similar results.

These results prove the robustness of the approximate and asymptotic design-unbiasedness of GREG estimators. This property is independent of the quality of the auxiliaries or the sign of the coefficient in the adjustment term of the GREG estimator.

We also investigated the consequences of applying integrated weights in regressions. To analyze this issue, we run weighted regressions with sex, age, and ms as defined in Table 3.6 as independent variables. Because how well the independent variables explain the dependent variable is not of interest, we included the same independent variables in all estimators to ensure comparability. The issue is rather whether it matters if equal weights are used for all persons within a household when predicting person-level characteristics. Table A.6 summarizes the dependent variables of different types.

Table A.6: Dependent variables

Variable	Description
empl_inc	Employee cash or near-cash income
unemp_bene	Unemployment benefits
sick_bene	Sickness benefits
mana_pos	Managerial position with two categories (supervisory responsible, non-supervisory responsible)

Figure A.1 depicts the RB of the coefficients of the weighted regressions. It becomes apparent that no considerable differences arise when individual GREG weights or equal integrated weights are applied. We used different scales in the table because the ranges of sex1 and mana_pos are large. Also, the number of MC replications with p-values < 0.05 are more or less equal for both integrated GREG estimators and a naïve GREG estimator (Table A.7 in Appendix A). The results shown in Figure A.1 and Table A.7, disprove that using integrated weights for regression analysis has considerable consequences.

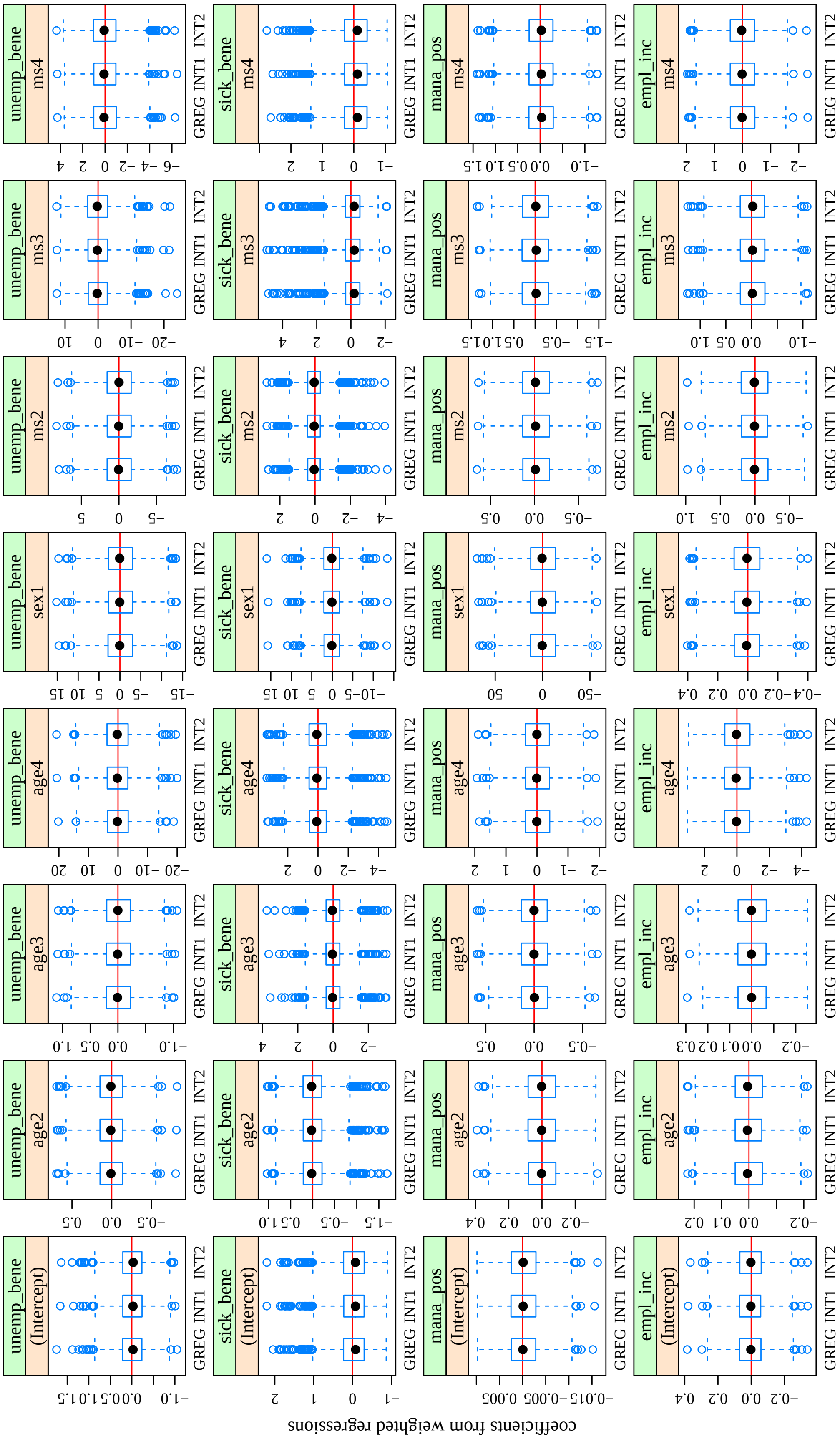

Figure A.1: Relative bias of $\hat{B}_r$ from weighted regressions for $m = 1500$

To conclude, even when the relationship between the variables of interest and the auxiliaries is reversed from the person to household level, we did not succeed in proving a bias for the integrated GREG estimator, which shows the robustness of the asymptotic design-unbiasedness of GREG estimators. In addition, the regression results are not influenced when equal integrated weights are used instead of individual weights.

Table A.7: Number MC replications with pvalue < 0.05 out of $R = 1,000$ for different variables of interest

		Intercept	age2	age3	age4	sex1	ms2	ms3	ms4
empl_inc	GREG	782	807	550	365	60	389	430	864
	INT1	785	803	548	361	58	383	426	859
	INT2	787	804	549	362	54	383	427	860
unempl_bene	GREG	1000	1000	1000	67	1000	977	883	334
	INT1	1000	1000	1000	66	1000	981	878	335
	INT2	1000	1000	1000	70	1000	981	873	334
sick_bene	GREG	544	995	853	22	61	114	49	75
	INT1	548	995	852	21	61	117	54	75
	INT2	549	996	852	22	60	117	54	73
mana_pos	GREG	1000	1000	999	251	44	997	651	714
	INT1	1000	1000	998	250	49	998	645	708
	INT2	1000	1000	999	255	49	997	645	711

B Additional Material for Chapter 4

B.1 Variance Estimator for the Second Weighting Approach

The following result describes the variance estimator for the second proposed weighting approach. Its proof is analogously given to the proof of Result 5.

Result 12. ***Variance Estimator for the Second Proposed Weighting Approach*** *The variance estimator of the second proposed person-level estimator $\hat{T}_{y_p}^{WA2}$ using the Taylor linearization technique is given by*

$$\begin{aligned}
\hat{V}_1 &= \hat{V}(\hat{T}_{y_p}^{GREG}), & \hat{V}_{12} &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^T \widehat{Cov}(\hat{T}_{y_p}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^*}^{GREG}),\\
\hat{V}_2 &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^*}^{GREG})\hat{\boldsymbol{D}}_{\boldsymbol{c}}, & \hat{V}_{13} &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^T \widehat{Cov}(\hat{T}_{y_p}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{GREG}),\\
\hat{V}_3 &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{GREG})\hat{\boldsymbol{D}}_{\boldsymbol{c}}, & \hat{V}_{23} &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^T \widehat{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^*}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{GREG})\hat{\boldsymbol{D}}_{\boldsymbol{c}}.
\end{aligned}$$

At the household level, the variance components are obtained from

$$\begin{aligned}
\hat{V}_1 &= \hat{V}(\hat{T}_{y_h}^{GREG}), & \hat{V}_{12} &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T \widehat{Cov}(\hat{T}_{y_h}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^*}^{GREG}),\\
\hat{V}_2 &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^*}^{GREG})\hat{\boldsymbol{E}}_{\boldsymbol{c}}, & \hat{V}_{13} &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T \widehat{Cov}(\hat{T}_{y_h}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{h}}}^{GREG}),\\
\hat{V}_3 &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{h}}}^{GREG})\hat{\boldsymbol{E}}_{\boldsymbol{c}}, & \hat{V}_{23} &= \hat{\boldsymbol{E}}_{\boldsymbol{c}}{}^T \widehat{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^*}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{h}}}^{GREG})\hat{\boldsymbol{E}}_{\boldsymbol{c}}.
\end{aligned}$$

$\widehat{Cov}$ denotes the estimated covariance. Estimated variances and covariances can be obtained by (2.10) by inserting the appropriate variables.

Proof. The linearized second proposed estimator at person-level is given by

$$\hat{T}_{y_p}^{\text{WA2}} \doteq \hat{T}_{y_p}^{\text{GREG}} + \boldsymbol{D}_{\boldsymbol{c}}{}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}^*}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c}_{\boldsymbol{p}}}^{\text{GREG}}).$$

Its variance is derived by

$$\begin{aligned}
\mathrm{V}(\hat{T}_{y_p}^{\mathrm{WA2}}) &\doteq \mathrm{V}(\hat{T}_{y_p}^{\mathrm{GREG}} + \boldsymbol{D_c}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}})) \\
&= \mathrm{V}(\hat{T}_{y_p}^{\mathrm{GREG}}) + \mathrm{V}(\boldsymbol{D_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}} - \boldsymbol{D_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}}) \\
&\quad + 2\mathrm{Cov}(\hat{T}_{y_p}^{\mathrm{GREG}}, \boldsymbol{D_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}} - \boldsymbol{D_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}}) \\
&= \mathrm{V}(\hat{T}_{y_p}^{\mathrm{GREG}}) + \boldsymbol{D_c}^T\mathrm{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}})\boldsymbol{D_c} + \boldsymbol{D_c}^T\mathrm{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}})\boldsymbol{D_c} \\
&\quad - 2\boldsymbol{D_c}^T\mathrm{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}})\boldsymbol{D_c} \\
&\quad + 2\boldsymbol{D_c}^T\mathrm{Cov}(\hat{T}_{y_p}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}}) - 2\boldsymbol{D_c}^T\mathrm{Cov}(\hat{T}_{y_p}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\mathrm{GREG}})
\end{aligned}$$

with Cov as approximate covariance. $\hat{V}(\hat{T}_{y_p}^{\mathrm{WA2}})$ results by estimating $\mathrm{V}(\hat{T}_{y_p}^{\mathrm{WA2}})$ from the sample s_p by the plug-in method. We continue with the household-level proposed estimator which is linearized by

$$\hat{T}_{y_h}^{\mathrm{WA2}} \doteq \hat{T}_{y_h}^{\mathrm{GREG}} + \boldsymbol{E_c}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}).$$

Its variance is derived by

$$\begin{aligned}
\mathrm{V}(\hat{T}_{y_h}^{\mathrm{WA2}}) &\doteq \mathrm{V}(\hat{T}_{y_h}^{\mathrm{GREG}} + \boldsymbol{E_c}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}})) \\
&= \mathrm{V}(\hat{T}_{y_h}^{\mathrm{GREG}}) + \mathrm{V}(\boldsymbol{E_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}} - \boldsymbol{E_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}) \\
&\quad + 2\mathrm{Cov}(\hat{T}_{y_h}^{\mathrm{GREG}}, \boldsymbol{E_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}} - \boldsymbol{E_c}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}) \\
&= \mathrm{V}(\hat{T}_{y_h}^{\mathrm{GREG}}) + \boldsymbol{E_c}^T\mathrm{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}})\boldsymbol{E_c} + \boldsymbol{E_c}^T\mathrm{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}})\boldsymbol{E_c} \\
&\quad - 2\boldsymbol{E_c}^T\mathrm{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}})\boldsymbol{E_c} \\
&\quad + 2\boldsymbol{E_c}^T\mathrm{Cov}(\hat{T}_{y_h}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p^*}}^{\mathrm{GREG}}) - 2\boldsymbol{E_c}^T\mathrm{Cov}(\hat{T}_{y_h}^{\mathrm{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\mathrm{GREG}}).
\end{aligned}$$

$\hat{V}(\hat{T}_{y_h}^{\mathrm{WA2}})$ results by estimating $\mathrm{V}(\hat{T}_{y_h}^{\mathrm{WA2}})$ from the sample s_h by the plug-in method. □

B.2 Composite Estimator given by Renssen and Nieuwenbroek

Renssen and Nieuwenbroek (1997) introduced the composite GREG estimator to align the estimates for variables common to two independent surveys. To estimate to unknown totals of the common variables they suggest using a weighted average of the estimates obtained from each of the independent surveys. Adopting this method to household surveys yields the following person- and household-level estimators

$$\begin{aligned} \hat{T}_{y_p}^{\text{RN}} &= \hat{T}_{y_p}^{\text{GREG}} + \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}(\tilde{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{RN}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}) \text{ , and} \\ \hat{T}_{y_h}^{\text{RN}} &= \hat{T}_{y_h}^{\text{GREG}} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}(\tilde{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{RN}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}), \end{aligned} \tag{B.1}$$

where $\hat{T}_{y_p}^{\text{GREG}}$, $\hat{T}_{y_h}^{\text{GREG}}$, $\hat{\boldsymbol{D}}_c$ and $\hat{\boldsymbol{E}}_c$ are defined as in (4.24) and (4.9), respectively.

Renssen and Nieuwenbroek (1997, p.371) suggested a composite estimator for the unknown population total $\tilde{\boldsymbol{T}}_c$ based on the weighted average of the single estimates obtained from each of the independent survey. Their suggested estimator was given by

$$\tilde{\boldsymbol{T}}_{\boldsymbol{c}}^{\text{RN}} = \boldsymbol{Q}\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} + (\boldsymbol{1} - \boldsymbol{Q})\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}, \tag{B.2}$$

where $\boldsymbol{Q}$ is a weighting matrix of dimension $(L \times L)$ with $\boldsymbol{Q} + (\boldsymbol{1} - \boldsymbol{Q}) = \boldsymbol{I}$. An optimal choice in the sense of minimizing the variance of the composite estimator $\boldsymbol{u}^T\tilde{\boldsymbol{T}}_c$ for an arbitrary L-vector $\boldsymbol{u}$ and considering the dependence between the person and the household data set is

$$\boldsymbol{Q} = [V(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) - \text{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})][V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}) + V(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) - 2\text{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})]^{-1}$$

with $V(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}})$ and $\text{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})$ as variance and covariance. As the variance and covariance are unknown, $\boldsymbol{Q}$ is replaced by its estimate.

Inserting (B.2) into (B.1) yields

$$\begin{aligned} \hat{T}_{y_p}^{\text{RN}} &= \hat{T}_{y_p}^{\text{GREG}} - \hat{\boldsymbol{D}}_{\boldsymbol{c}}^{T}(\boldsymbol{1} - \boldsymbol{Q})(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \text{ , and} \\ \hat{T}_{y_h}^{\text{RN}} &= \hat{T}_{y_h}^{\text{GREG}} + \hat{\boldsymbol{E}}_{\boldsymbol{c}}^{T}\boldsymbol{Q}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}). \end{aligned} \tag{B.3}$$

It can be seen that the higher the difference between the person- and the household-level estimate of the common variables, the higher the adjustment term. The resulting weights from (B.3) are obtained by

$$\begin{aligned} w_i^{\text{RN}} &= w_i^{\text{GREG}} - \sum_{i \in s_p} \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}\,T}(\sum_{i \in s_p} \boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}}\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{F_x}\,T})^{-1}(\boldsymbol{1} - \boldsymbol{Q})(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \text{ , and} \\ w_g^{\text{RN}} &= w_g^{\text{GREG}} + \sum_{g \in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}\,T}(\sum_{g \in s_h} \boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}}\boldsymbol{r}_{\boldsymbol{g}}^{\boldsymbol{F_a}\,T})^{-1}\boldsymbol{Q}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}). \end{aligned}$$

The estimated variances of the person-level estimator using the Taylor linearization technique is given by (Renssen and Nieuwenbroek, 1997, p.371)

$$\hat{V}(\hat{T}^{\mathrm{RN}}_{y_p}) \doteq \widehat{V}_1 + \widehat{V}_2 + \widehat{V}_3 - 2\widehat{\mathrm{V}}_{12} - 2\widehat{\mathrm{V}}_{13} + 2\widehat{\mathrm{V}}_{23}$$

with

$$\begin{aligned}
\widehat{V}_1 &= \hat{V}(\hat{T}^{\mathrm{GREG}}_{y_h}),\\
\widehat{V}_2 &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^{T}(\mathbf{1}-\boldsymbol{Q})\hat{V}(\hat{\boldsymbol{T}}^{\mathrm{GREG}}_{\boldsymbol{c_p}})(\mathbf{1}-\boldsymbol{Q})^{T}\hat{\boldsymbol{D}}_{\boldsymbol{c}},\\
\widehat{V}_3 &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^{T}(\mathbf{1}-\boldsymbol{Q})\hat{V}(\hat{\boldsymbol{T}}^{\mathrm{GREG}}_{\boldsymbol{c_h}})(\mathbf{1}-\boldsymbol{Q})^{T}\hat{\boldsymbol{D}}_{\boldsymbol{c}},\\
\widehat{\mathrm{V}}_{12} &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^{T}(\mathbf{1}-\boldsymbol{Q})\widehat{\mathrm{Cov}}(\hat{T}^{\mathrm{GREG}}_{y_h},\hat{\boldsymbol{T}}^{\mathrm{GREG}}_{\boldsymbol{c_p}})\\
\widehat{\mathrm{V}}_{13} &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^{T}(\mathbf{1}-\boldsymbol{Q})\widehat{\mathrm{Cov}}(\hat{T}^{GREG}_{y_h},\hat{\boldsymbol{T}}^{\mathrm{GREG}}_{\boldsymbol{c_h}}),\\
\widehat{\mathrm{V}}_{23} &= \hat{\boldsymbol{D}}_{\boldsymbol{c}}{}^{T}(\mathbf{1}-\boldsymbol{Q})\widehat{\mathrm{Cov}}(\hat{\boldsymbol{T}}^{\mathrm{GREG}}_{\boldsymbol{c_p}},\hat{\boldsymbol{T}}^{\mathrm{GREG}}_{\boldsymbol{c_h}})(\mathbf{1}-\boldsymbol{Q})^{T}\hat{\boldsymbol{D}}_{\boldsymbol{c}}.
\end{aligned}$$

Estimated variances and covariances can be obtained by (2.10) by inserting the appropriate variables. The variance estimator of the household-level estimator is given in a similar manner.

B.3 Variance Estimator for the GLS Estimator

The following result describes the variance estimators via Taylor linearization for the GLS estimator given by Zieschang (1986, 1990).

Result 13. ***Variance Estimators for the GLS Estimator***
The variance estimator for the combined calibration estimator at person-level (4.35) *using the Taylor linearization technique is given by*

$$\hat{V}(\hat{T}_{y_p}^{ZIE}) \doteq \hat{V}_1 + \hat{V}_2 + \hat{V}_3 - 2\hat{V}_{12} + 2\hat{V}_{13} - 2\hat{V}_{23}$$

with

$$\begin{aligned} \hat{V}_1 &= \hat{V}(\hat{T}_{y_p}^{GREG}), & \hat{V}_{12} &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T \widehat{Cov}(\hat{T}_{y_p}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}), \\ \hat{V}_2 &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG})\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}, & \hat{V}_{13} &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T \widehat{Cov}(\hat{T}_{y_p}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}), \\ \hat{V}_3 &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG})\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}, & \hat{V}_{23} &= \hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}{}^T \widehat{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG})\hat{\boldsymbol{D}}_{\boldsymbol{\kappa}}. \end{aligned}$$

At household-level the variance estimator of the combined calibration estimator (4.36) *using the Taylor linearization technique is obtained from*

$$\hat{V}(\hat{T}_{y_h}^{ZIE}) \doteq \hat{V}_1 + \hat{V}_2 + \hat{V}_3 + 2\hat{V}_{12} - 2\hat{V}_{13} - 2\hat{V}_{23}$$

with

$$\begin{aligned} \hat{V}_1 &= \hat{V}(\hat{T}_{y_h}^{GREG}), & \hat{V}_{12} &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^T \widehat{Cov}(\hat{T}_{y_h}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}), \\ \hat{V}_2 &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG})\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}, & \hat{V}_{13} &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^T \widehat{Cov}(\hat{T}_{y_h}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG}), \\ \hat{V}_3 &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^T \hat{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG})\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}, & \hat{V}_{23} &= \hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}{}^T \widehat{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{GREG}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{GREG})\hat{\boldsymbol{E}}_{\boldsymbol{\kappa}}. \end{aligned}$$

$\widehat{Cov}$ *denotes the estimated covariance. Estimated variances and covariances can be obtained by* (2.10) *by inserting the appropriate variables.*

Proof. Analogously to the proof of Result 5, the Taylor linarization technique for the person-level estimator yields

$$\hat{T}_{y_p}^{\text{ZIE}} \doteq \hat{T}_{y_p}^{\text{GREG}} - \boldsymbol{D}_{\boldsymbol{\kappa}}{}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}).$$

Its variance is derived by

$$\begin{aligned} \mathrm{V}(\hat{T}_{y_p}^{\text{ZIE}}) &\doteq \mathrm{V}(\hat{T}_{y_p}^{\text{GREG}} - \boldsymbol{D}_{\boldsymbol{\kappa}}{}^T(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})) \\ &= \mathrm{V}(\hat{T}_{y_p}^{\text{GREG}}) + \mathrm{V}(\boldsymbol{D}_{\boldsymbol{\kappa}}{}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \boldsymbol{D}_{\boldsymbol{\kappa}}{}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \\ &\qquad - 2\mathrm{Cov}(\hat{T}_{y_p}^{\text{GREG}}, \boldsymbol{D}_{\boldsymbol{\kappa}}{}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}} - \boldsymbol{D}_{\boldsymbol{\kappa}}{}^T\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \\ &= \mathrm{V}(\hat{T}_{y_p}^{\text{GREG}}) + \boldsymbol{D}_{\boldsymbol{\kappa}}{}^T\mathrm{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}})\boldsymbol{D}_{\boldsymbol{\kappa}} + \boldsymbol{D}_{\boldsymbol{\kappa}}{}^T\mathrm{V}(\hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})\boldsymbol{D}_{\boldsymbol{\kappa}}{}^T \\ &\qquad - 2\boldsymbol{D}_{\boldsymbol{\kappa}}{}^T\mathrm{Cov}(\hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}})\boldsymbol{D}_{\boldsymbol{\kappa}}{}^T \\ &\qquad - 2\boldsymbol{D}_{\boldsymbol{\kappa}}{}^T\mathrm{Cov}(\hat{T}_{y_p}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_p}}^{\text{GREG}}) + 2\boldsymbol{D}_{\boldsymbol{\kappa}}{}^T\mathrm{Cov}(\hat{T}_{y_p}^{\text{GREG}}, \hat{\boldsymbol{T}}_{\boldsymbol{c_h}}^{\text{GREG}}) \end{aligned}$$

with Cov as approximate covariance. $\hat{V}(\hat{T}^{\text{ZIE}}_{y_p})$ results by estimating $\mathrm{V}(\hat{T}^{\text{ZIE}}_{y_p})$ from the sample s_p by the plug-in method. We continue with the household-level proposed estimator which is linearized by

$$\hat{T}^{\text{ZIE}}_{y_h} \doteq \hat{T}^{\text{GREG}}_{y_h} + \boldsymbol{E_c}^T(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c}^*_{\boldsymbol{p}}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}}).$$

Its variance is derived by

$$\begin{aligned}
\mathrm{AV}(\hat{T}^{\text{ZIE}}_{y_h}) &\doteq \mathrm{V}(\hat{T}^{\text{GREG}}_{y_h} + \boldsymbol{E_c}^T(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})) \\
&= \mathrm{V}(\hat{T}^{\text{GREG}}_{y_h}) + \mathrm{V}(\boldsymbol{E_c}^T\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \boldsymbol{E_c}^T\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}}) \\
&\quad + 2\mathrm{Cov}(\hat{T}^{\text{GREG}}_{y_h}, \boldsymbol{E_c}^T\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}} - \boldsymbol{E_c}^T\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}}) \\
&= \mathrm{V}(\hat{T}^{\text{GREG}}_{y_h}) + \boldsymbol{E_c}^T\mathrm{V}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c}^*_{\boldsymbol{p}}})\boldsymbol{E_c} + \boldsymbol{E_c}^T\mathrm{V}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})\boldsymbol{E_c} \\
&\quad - 2\boldsymbol{E_c}^T\mathrm{Cov}(\hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}})\boldsymbol{E_c} \\
&\quad + 2\boldsymbol{E_c}^T\mathrm{Cov}(\hat{T}^{\text{GREG}}_{y_h}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_p}}) - 2\boldsymbol{E_c}^T\mathrm{Cov}(\hat{T}^{\text{GREG}}_{y_h}, \hat{\boldsymbol{T}}^{\text{GREG}}_{\boldsymbol{c_h}}).
\end{aligned}$$

$\hat{V}(\hat{T}^{\text{ZIE}}_{y_h})$ results by estimating $\mathrm{V}(\hat{T}^{\text{ZIE}}_{y_h})$ from the sample s_h by the plug-in method. □

B.4 Specialized Auxiliary Variable Sets to Estimate the Unknown Common Variable Totals

Table B.1: Specialized auxiliary variable set to estimate `inc`

Variable name in AMELIA	Description
AGE	Age with four categories
SEX	Sex with two categories
MST	Marital status with three categories
BAS	Basic activity status with four categories
HHS	Household size
PY090	Unemployment benefits

Table B.2: Specialized auxiliary variable set to estimate `soc`

Variable name in AMELIA	Description
AGE	Age with four categories
SEX	Sex with two categories
MST	Marital status with three categories
BAS	Basic activity status with four categories
HHS	Household size
PY090	Unemployment benefits
PY010	Employee cash or near-cash income
SEM	Self-employment dummy

B.5 Additional Tables for the Simulation Study

	m=1500						m=200					
	INT1	INT2	WA1	WA2	ZIE	MER	INT1	INT2	WA1	WA2	ZIE	MER
`inc`	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00
`soc`	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00
`sel`	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00
`act1`	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
`act2`	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00
`act3`	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
`inc_hs1`	-0.01	-0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.00	0.00	-0.00	0.00	0.00
`inc_hs2`	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.01	0.00	-0.00	-0.00	-0.00	-0.00
`inc_hs3`	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00
`inc_hs4`	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00
`inc_hs5`	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
`inc_hs6`	-0.01	-0.01	-0.00	-0.00	-0.00	-0.00	-0.01	-0.01	-0.00	-0.01	-0.01	-0.01
`bene_age1`	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	-0.00	0.01	0.01
`bene_age2`	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.00	0.02	0.02
`bene_age3`	0.00	0.00	0.00	-0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.02	0.02
`bene_age4`	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01

the person level

mates at the household level

	m=1500						m=200					
	INT1	INT2	WA1	WA2	ZIE	MER	INT1	INT2	WA1	WA2	ZIE	MER
inc	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00
soc	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00
gross_inc	0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00
cap_inc	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00
taxes	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00

C Additional Material for Chapter 5

C.1 Theorems and Proofs Given by Steel and Clark (2007)

In this section, we deeply discuss the theorems and proofs given by Steel and Clark (2007). When enriching the comprehension we declare skipped intermediate calculations. Original text from Steel and Clark (2007) is indicated by boxes. For a better understanding, we change the original notation into the notation of the present thesis.

C.1.1 Theorem 1: Optimal Estimator for Simple Cluster Sampling

First theorem given by Steel and Clark (2007, p. 53): Optimal estimator for simple cluster sampling

Suppose that m households are selected by simple random sampling without replacement from a population of M households, and all people are selected from selected households. Consider the estimator of T_y given by

$$\hat{T}_y = \hat{T}_y^{\text{HT}} + \boldsymbol{h}^T(\boldsymbol{T_x} - \boldsymbol{\hat{T}_x^{\text{HT}}}),$$

where $\boldsymbol{h}$ is a constant Q-vector. It is assumed that there exists a vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^T\boldsymbol{x_i} = 1$ for all $i \in U$. The variance of this estimator is minimized by $\boldsymbol{h}^*$ which are solutions of

$$\sum_{g \in s_h} (y_g - \boldsymbol{h}^T\boldsymbol{x_g})\boldsymbol{x_g} = \boldsymbol{0}. \tag{C.1}$$

Hence $\hat{T}_y^{\text{INT}}$ with $v_i = N_g^{-1}$ for all $i \in U_p$ is the optimal choice of $\hat{T}_y$.

C.1.1.1 Proof of Theorem 1:

Let $\bar{Y} = \frac{T_y}{M}$ and $\bar{\boldsymbol{X}} = \frac{\boldsymbol{T_x}}{M}$ be the population means of y_g and $\boldsymbol{x_g}$. The variance of $\hat{T}_y$ is

$$\begin{aligned} V(\hat{T}_y) &= V\left[\widehat{T}_y^{\mathrm{HT}} + \boldsymbol{h}^T(\boldsymbol{T}_x - \widehat{\boldsymbol{T}}_x^{\mathrm{HT}})\right] \\ &= V\left(\frac{M}{m}\sum_{s_h}(y_g - \boldsymbol{h}^T\boldsymbol{x}_g)\right), \\ &= \frac{M^2}{m}\left(1 - \frac{m}{M}\right)S_r^2 \end{aligned} \tag{C.2}$$

where $S_r^2 = (M-1)^{-1}\sum_{g\in U_h}\left(y_g - \boldsymbol{h}^T\boldsymbol{x_g} - (\bar{Y}_I - \boldsymbol{h}^T\bar{\boldsymbol{X}})\right)^2$.

Since the proof is based on large sample properties in general and in particular also the formula of S_r^2 refers to the population $\sum_{U_h}$ it should be $\sum_{U_h}$, not $\sum_{s_h}$.

To minimize with respect to **h**, we set the derivative of S_r^2 to zero

$$\begin{aligned} 0 &= (M-1)^{-1}\sum_{g\in U_h}\left(y_g - \boldsymbol{h}^T\boldsymbol{x_g} - (\bar{Y} - \boldsymbol{h}^T\bar{\boldsymbol{X}})\right)(\boldsymbol{x_g} - \bar{\boldsymbol{X}}) \\ &= \sum_{g\in U_h}\left(y_g - \boldsymbol{h}^T\boldsymbol{x_g} - (\bar{Y} - \boldsymbol{h}^T\bar{\boldsymbol{X}})\right)\boldsymbol{x_g} - \sum_{g\in U_h}\left((y_g - \bar{Y}) - \boldsymbol{h}^T(\boldsymbol{x_g} - \bar{\boldsymbol{X}}_I\right)\bar{\boldsymbol{X}}_I \\ &= \sum_{g\in U_h}\left(y_g - \boldsymbol{h}^T\boldsymbol{x_g} - (\overline{Y} - \boldsymbol{h}^T\bar{\boldsymbol{X}})\right)\boldsymbol{x_g} \\ &= \sum_{g\in U_h}\left(y_g - \boldsymbol{h}^T\boldsymbol{x_g}\right)\boldsymbol{x_g} - (\bar{Y}_I - \boldsymbol{h}^T\bar{\boldsymbol{X}}_I)\boldsymbol{T_x}. \end{aligned} \tag{C.3}$$

0 should be declared as a vector.

For a better traceability, we add some skipped intermediate calculations

$$\begin{aligned} &\bar{\boldsymbol{X}}\sum_{g\in U_h} y_g - \bar{\boldsymbol{X}}\sum_{g\in U_h}\bar{Y} - \bar{\boldsymbol{X}}\boldsymbol{h}^T\sum_{g\in U_h}\boldsymbol{x_g} + \bar{\boldsymbol{X}}\boldsymbol{h}^T\sum_{g\in U_h}\bar{\boldsymbol{X}} \\ &\quad = \underbrace{\frac{\boldsymbol{T_x}}{M}T_y - \frac{\boldsymbol{T_x}}{M}M\frac{T_y}{M}}_{\boldsymbol{0}} - \underbrace{\frac{\boldsymbol{T_x}}{M}\boldsymbol{h}^T\boldsymbol{T_x} + \frac{\boldsymbol{T_x}}{M}M\boldsymbol{h}^T\frac{\boldsymbol{T_x}}{M}}_{\boldsymbol{0}} = \mathbf{0}. \end{aligned}$$

Now we show that (C.3) is satisfied by $\boldsymbol{h}^*$. By assumption, $\boldsymbol{h}^*$ satisfies

$$\boldsymbol{0} = \sum_{g \in U_I} (y_g - {\boldsymbol{x_g}}^T \boldsymbol{h}^*) \boldsymbol{x_g}. \tag{C.4}$$

Hence, the first sum in the right hand side of (C.3) is equal to zero for $\boldsymbol{h} = \boldsymbol{h}^*$. Premultiplying both sides of (C.4) by $\boldsymbol{\lambda}^T$ gives

$$\begin{aligned} 0 &= \sum_{g \in U_I} (y_g - {\boldsymbol{x_g}}^T \boldsymbol{h}^*) \boldsymbol{\lambda}^T \boldsymbol{x_g} \\ 0 &= \sum_{g \in U_I} (y_g - {\boldsymbol{x_g}}^T \boldsymbol{h}^*) \\ 0 &= T_y - {\boldsymbol{T_x}}^T \boldsymbol{h}^*. \end{aligned} \tag{C.5}$$

Dividing by M gives $\bar{Y}_I - {\bar{\boldsymbol{X}}_I}^T \boldsymbol{h}^* = 0$. Hence, the rest of the right hand side of (C.3) is equal to zero. So $\boldsymbol{h}^*$ satisfies (C.3).

However, following their statement in Theorem 1 assuming $\boldsymbol{\lambda}^T \boldsymbol{x}_i = 1$, the last paragraph of the proof of Theorem 1 should be changed to

$$\begin{aligned} 0 &= \sum_{g \in U_h} (y_g - {\boldsymbol{x_g}}^T \boldsymbol{h}^*) \boldsymbol{\lambda}^T \boldsymbol{x_g} \\ 0 &= \sum_{g \in U_h} (y_g - {\boldsymbol{x_g}}^T \boldsymbol{h}^*) N_g \\ 0 &= N\bar{Y} - N{\bar{\boldsymbol{X}}}^T \boldsymbol{h}^*, \end{aligned}$$

because it is valid that

$$\boldsymbol{\lambda}^T \boldsymbol{x_g} = \boldsymbol{\lambda}^T \sum_{i \in U_g} \boldsymbol{x_i} = \sum_{i \in U_g} \boldsymbol{\lambda}^T \boldsymbol{x_i} = \sum_{i \in U_g} 1 = N_g.$$

To ensure that the calculations of Steel and Clark (2007) are correct, $\boldsymbol{\lambda}^T \boldsymbol{x_i} = N_g$ should be assumed instead of $\boldsymbol{\lambda}^T \boldsymbol{x_i} = 1$. Nevertheless, their result is not affected by the wrong assumption because $N\bar{Y} - N{\bar{\boldsymbol{X}}}^T \boldsymbol{h}^*$ still equals zero.

C.1.2 Theorem 2: Explaining the Difference in the Asymptotic Variances

Second theorem given by Steel and Clark (2007, p.54): Explaining the difference in the asymptotic variances

Suppose that m households are selected by simple random sampling without replacement and all people are selected from selected households. Let $r_i^{B_p} = y_i - \boldsymbol{B_p}^T \boldsymbol{x_i}$ and let $\boldsymbol{B_c}$ be the result of regressing $r_i^{B_p}$ on $\boldsymbol{\bar{x}_g}$ over $i \in U_p$ using weighted least squares regression weighted by N_g. Then

$$V(\hat{T}_y^{\text{GREG}}) - V(\hat{T}_y^{\text{INT}}) = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1}\, \boldsymbol{B_c}^T \sum_{g \in U_h} \left(\boldsymbol{x_g} \boldsymbol{x_g}^T\right) \boldsymbol{B_c} \tag{C.6}$$

where $\hat{T}_y^{\text{INT}}$ is calculated using $v_i = N_g^{-1}$ for all $i \in U_p$.

C.1.2.1 Proof of Theorem 2:

Let “-“ denote a generalized inverse of a matrix. Then $\boldsymbol{B_c}$ is equal to

$$\begin{aligned}\boldsymbol{B_c} &= \left(\sum_{g \in U_h} \sum_{i \in U_g} N_g \boldsymbol{\bar{x}_g} \boldsymbol{\bar{x}_g}^T\right)^{-} \sum_{g \in U_h} \sum_{i \in U_g} N_g \boldsymbol{\bar{x}_g} r_i \\ &= \left(\sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{x_g}^T\right)^{-} \sum_{g \in U_h} \boldsymbol{x_g} r_g. \end{aligned} \tag{C.7}$$

Now, $r_i = y_i - \boldsymbol{B_p}^T \boldsymbol{x_i}$ so $r_g = y_g - \boldsymbol{B_p}^T \boldsymbol{x_g}$.

As a remark, the line is derived by $\sum_{i \in U_g} r_i = \sum_{i \in U_g} (y_i - \boldsymbol{B_p}^T \boldsymbol{x_i}) = \sum_{i \in U_g} y_i - \boldsymbol{B_p}^T \sum_{i \in U_g} \boldsymbol{x_i} = y_g - \boldsymbol{B_p}^T \boldsymbol{x_g} = r_g$. Nevertheless, the aggregation does not imply that the aggregated residuals per household r_g equal the residuals from a household-level regression: $y_g = \boldsymbol{B_h}^T \boldsymbol{x_g} + r_g^{B_h}$ with $\boldsymbol{B_p} \neq \boldsymbol{B_h}$ as well as $r_g \neq r_g^{B_h}$.

Hence, (C.7) becomes

$$\begin{aligned}\boldsymbol{B_c} &= \left(\sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{x_g}^T\right)^{-} \sum_{g \in U_h} \boldsymbol{x_g} \left(y_g - \boldsymbol{B_p}^T \boldsymbol{x_g}\right) \\ &= \left(\sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{x_g}^T\right)^{-} \sum_{g \in U_I} \boldsymbol{x_g} y_g - \left(\sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{x_g}^T\right)^{-} \sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{x_g}^T \boldsymbol{B_p} \\ &= \boldsymbol{B_h} - \boldsymbol{B_p} \end{aligned} \tag{C.8}$$

since $\boldsymbol{B_h} = \left(\sum_{g \in U_h} \boldsymbol{x_g} \boldsymbol{x_g}^T\right) \sum_{g \in U_h} \sum_{i \in U_g} \boldsymbol{x_g} y_g$.

The resulting formula should be $\boldsymbol{B_h} = \left(\sum_{g\in U_h} \boldsymbol{x_g}\boldsymbol{x_g}^T\right)^- \sum_{g\in U_h} \boldsymbol{x_g} y_g$.

The difference in the variances is given by

$$\begin{aligned} V(\hat{T}_y^{\text{GREG}}) - V(\hat{T}_y^{\text{INT}}) =& \frac{M^2}{m}\left(1 - \frac{m}{M}\right) \\ & (M-1)^{-1}\Big(\sum_{g\in U_I} (y_g - \boldsymbol{B_p}^T\boldsymbol{x_g})^2 - \sum_{g\in U_h} (y_g - \boldsymbol{B_h}^T\boldsymbol{x_g})^2\Big) \end{aligned}$$

For a better traceability, we add some skipped intermediate calculations

$$\begin{aligned} V(\hat{T}_y^{\text{GREG}}) - V(\hat{T}_y^{\text{INT}}) &= \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1} \\ & \sum_{g\in U_h} \left(y_g - \boldsymbol{B_p}^T\boldsymbol{x_g} - (\bar{Y} - \boldsymbol{B_p}^T\bar{\boldsymbol{X}})\right)^2 - \sum_{g\in U_h} \left(y_g - \boldsymbol{B_h}^T\boldsymbol{x_g} - (\bar{Y} - \boldsymbol{B_h}^T\bar{\boldsymbol{X}})\right)^2 \end{aligned}$$

with $\bar{Y} - \boldsymbol{B_p}^T\bar{\boldsymbol{X}} = \bar{Y} - \boldsymbol{B_h}^T\bar{\boldsymbol{X}} = 0$.

which becomes

$$\begin{aligned} V(\hat{T}_y^{\text{GREG}}) - V(\hat{T}_y^{\text{INT}}) & \Big/ \frac{M^2}{m}(1 - \frac{m}{M})(M-1)^{-1} \\ &= \sum_{g\in U_h} r_g^2 - \sum_{g\in U_I} \left(r_g + \boldsymbol{B_p}^T\boldsymbol{x_g} - \boldsymbol{B_h}^T\boldsymbol{x_g}\right)^2 \\ &= \sum_{g\in U_h} r_g^2 - \sum_{g\in U_I} \left(r_g - \boldsymbol{B_c}^T\boldsymbol{x_g}\right)^2 \\ &= \sum_{g\in U_h} \left(r_g - \boldsymbol{B_c}^T\boldsymbol{x_g} + \boldsymbol{B_c}^T\boldsymbol{x_g}\right)^2 - \sum_{g\in U_h} (r_g - \boldsymbol{B_c}^T\boldsymbol{x_g})^2 \\ &= \sum_{g\in U_h} \left(r_g - \boldsymbol{B_c}^T\boldsymbol{x_g}\right)^2 + \sum_{g\in U_h} \left(\boldsymbol{B_c}^T\boldsymbol{x_g}\right)^2 + 2\sum_{g\in U_h} \left(r_g - \boldsymbol{B_c}^T\boldsymbol{x_g}\right)\boldsymbol{x_g}^T\boldsymbol{B_c}^T \\ &\quad - \sum_{g\in U_h} \left(r_g - \boldsymbol{B_c}^T\boldsymbol{x_g}\right)^2 \\ &= \sum_{g\in U_h} \boldsymbol{B_c}^T\boldsymbol{x_g}\boldsymbol{x_g}^T\boldsymbol{B_c} + 2\sum_{g\in U_I} \left(r_g - \boldsymbol{B_c}^T\boldsymbol{x_g}\right)\boldsymbol{x_g}^T\boldsymbol{B_c}. \end{aligned} \tag{C.9}$$

Now, $\boldsymbol{B_c}$ is an ordinary least squares regression of r_g on $\boldsymbol{x_g}$ so

$$\sum_{g\in U_h} \left(r_g - \boldsymbol{B_c}^T\boldsymbol{x_g}\right)\boldsymbol{x_g} = \boldsymbol{0}.$$

Hence, (C.9) becomes

$$V(\hat{T}_y^{\text{GREG}}) - V(\hat{T}_y^{\text{INT}}) = \frac{M^2}{m}\left(1 - \frac{m}{M}\right)(M-1)^{-1}\boldsymbol{B_c}^T \sum_{g\in U_h} \boldsymbol{x_g}\boldsymbol{x_g}^T\boldsymbol{B_c}. \tag{C.10}$$

C.2 Additional Graphs for the Simulation Study

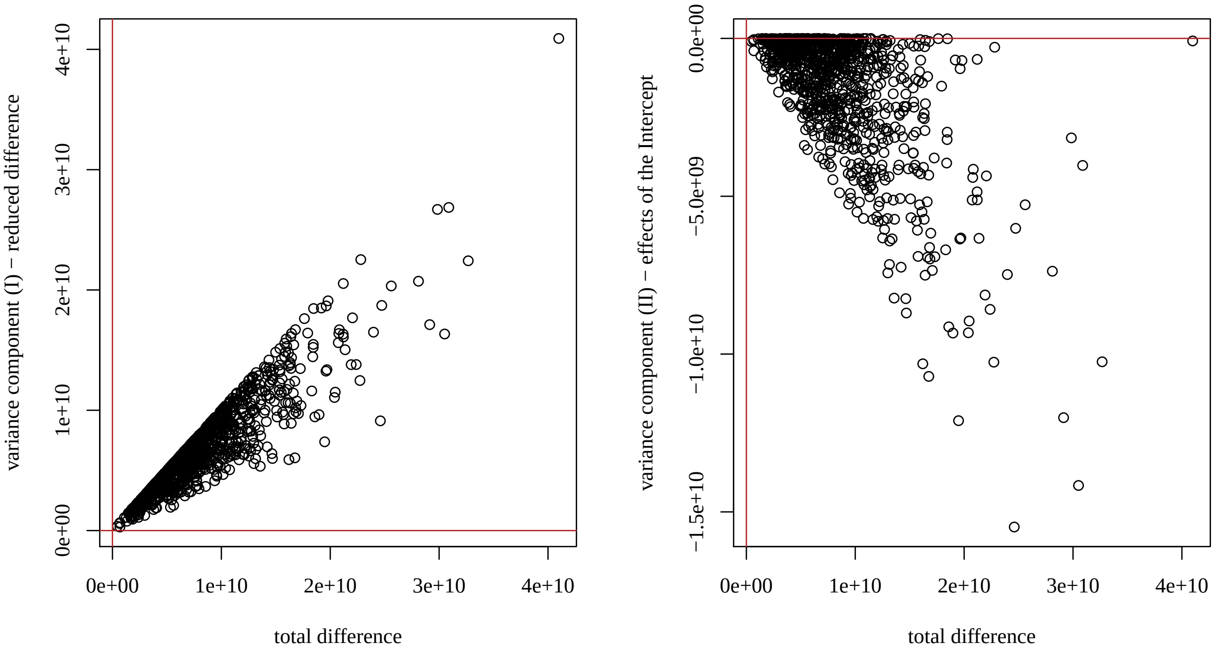

Figure C.1: Plots of the total difference against variances components I and II for case a) and $m = 1500$

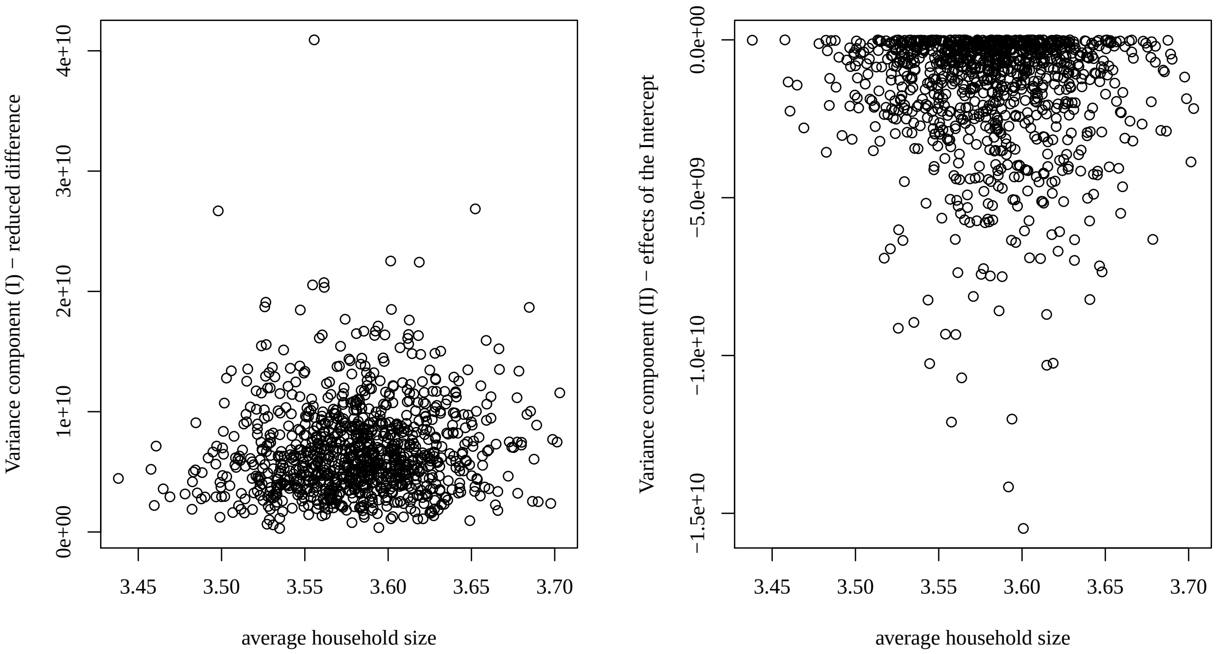

Figure C.2: Plots of the intercept and reduced difference against the average household size for case b) and $m = 1500$

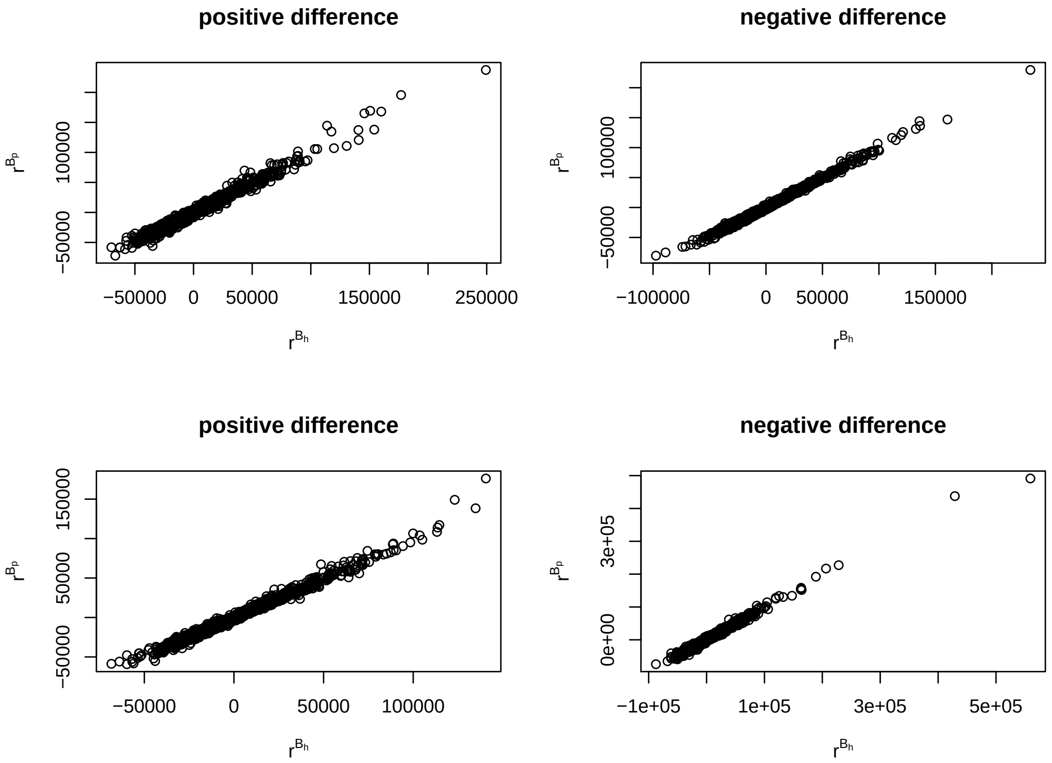

Figure C.3: Plots of the residuals for case b)

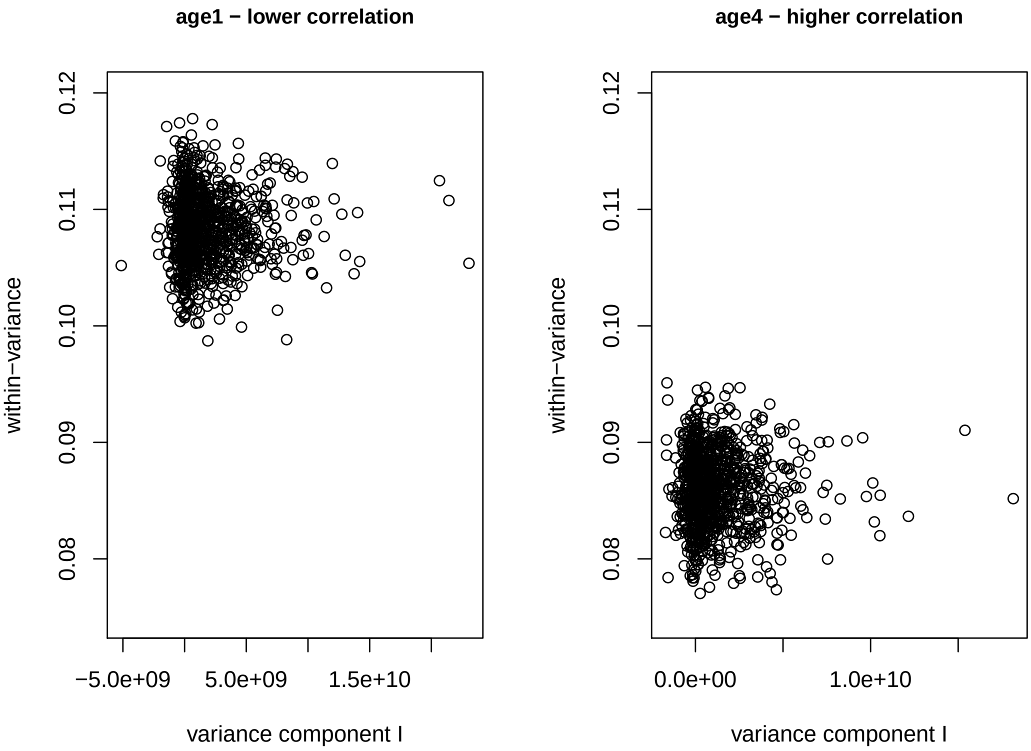

Figure C.4: Plots of the within-variance against variance component I for case b)

Bibliography

Afentakis, A. and Bihler, W. (2005). Das Hochrechnungsverfahren beim unterjährigen Mikrozensus ab 2005. *Wirtschaft und Statistik*, 10:1039–1048.

Alexander, C. H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13(2):183–198.

Antal, E. and Rothenbühler, M. (2015). Weighting in the Swiss household panel technical report. Technical report.

Backhaus, K., Erichson, B., Plinke, W., Schuchard-Ficher, C., and Weiber, R. (2008). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Berlin Heidelberg: Springer. doi:10.1007/978-3-662-56655-8.

Bankier, M. D. (1989). Generalized least squares estimation under poststratification. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 730–735.

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173. doi:10.1037//0022-3514.51.6.1173.

Basu, D. (2010). An essay on the logical foundations of survey sampling, part one. In *Selected Works of Debabrata Basu*, pages 167–206. Springer, New York. doi:10.1007/978-1-4419-5825-9_24.

Berger, Y. G., Muñoz, J. F., and Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals—an application to the extended regression estimator and the regression composite estimator. *Computational Statistics & Data Analysis*, 53(7):2596–2604. doi:10.1016/j.csda.2008.12.011.

Berger, Y. G., Tirari, M. E., and Tillé, Y. (2003). Towards optimal regression estimation in sample surveys. *Australian & New Zealand Journal of Statistics*, 45(3):319–329. doi:10.1111/1467-842x.00286.

Bethlehem, J., Cobben, F., and Schouten, B. (2011). *Handbook of nonresponse in household surveys*, volume 568. New York: John Wiley & Sons, Inc. doi:10.1002/9780470891056.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85(409):38–45. doi:10.2307/2289523.

Binder, D. A. (1982). Non-parametric bayesian models for samples from finite populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):388–393.

Boonstra, H., van den Brakel, J., Knotterus, P., Nieuwenbroek, N., and Renssen, R. (2003). A strategy to obtain consistency among tables of survey estimates. Technical report, DACSEIS report D7.2.

Branson, N. and Wittenberg, M. (2014). Reweighting South African national household survey data to create a consistent series over time: a cross-entropy estimation approach. *South African Journal of Economics*, 82(1):19–38. doi:10.1111/saje.12017.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4):831–846. doi:10.1093/biomet/92.4.831.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28(4):1026–1053.

Breidt, F. J. and Opsomer, J. D. (2009). Nonparametric and semiparametric estimation in complex surveys. In *Handbook of Statistics*, pages 103–119. Elsevier. doi:10.1016/s0169-7161(09)00227-2.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205. doi:10.1214/16-sts589.

Brewer, K. R. W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian & New Zealand Journal of Statistics*, 5(3):93–105. doi:10.1111/j.1467-842x.1963.tb00288.x.

Burgard, J. P., Kolb, J.-P., Merkle, H., and Münnich, R. (2017). Synthetic data for open and reproducible methodological research in social sciences and official statistics. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 11(3-4):233–244. doi:10.1007/s11943-017-0214-8.

Cassel, C.-M., Särndal, C., and Wretman, J. (1977). *Foundations of inference in survey sampling*. Wiley, New York. doi:10.2307/1403214.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620. doi:10.1093/biomet/63.3.615.

Chambers, R. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12(1):3.

Chambers, R. (2011). Which sample survey strategy? A review of three different approaches. *Centre for Statistical & Survey Methodology*, Working Paper(09-11):1–30.

Chen, J. and Sitter, R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9(2):385–406.

Clark, R. G. and Steel, D. G. (2002). The effect of using household as a sampling unit. *International Statistical Review*, 70(2):289–314. doi:10.2307/1403911.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons. doi:10.2307/2347176.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge. doi:10.4324/9781410606266.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444. doi:10.1214/aoms/1177731829.

Demnati, A. and Rao, J. (2004). Lineraization variance estimators for survey data. *Survey Methodology*, 30(1):17–26.

Deng, L.-Y. and Wu, C. F. J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82(398):568–576.

Destatis (2016). Bundesstatistikgesetz – BStatG. Retrieved from `https://www.destatis.de/DE/Methoden/Rechtsgrundlagen/Statistikbereiche/Inhalte/010_BStatG.pdf?__blob=publicationFile`. visited 25/07/2018.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382. doi:10.1080/01621459.1992.10475217.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, 88(423):1013–1020. doi:10.1080/01621459.1993.10476369.

Ericson, W. A. (1969). Subjective bayesian models in sampling finite populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(2):195–233.

Ericson, W. A. (1988). Bayesian inference in finite populations. In Krishnaiah, P. and Rao, C., editors, *Handbook of Statistics*, volume 6, pages 213–246. Elsevier Science Publishers, Amsterdam. doi:10.1016/s0169-7161(88)06011-0.

Estevao, V. M. and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16(4):379–399.

Estevao, V. M. and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2):127–147. doi:10.1111/j.1751-5823.2006.tb00165.x.

European Commission (2014). Methodological guidelines and description of EU-SILC target variables. Technical Report Doc-SILC065, Eurostat, Directorate F: Social Statistics.

Eurostat (2011). European statistics code of practice. Retrieved from `https://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15`. visited 23/10/2018.

Fahrmeir, L., Kneib, T., and Lang, S. (2007). *Regression*. Springer. doi:10.1007/978-3-642-01837-4.

Fairchild, A. J. and MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, 10(2):87–99. doi:10.1007/s11121-008-0109-6.

Firth, D. and Bennett, K. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):3–21. doi:10.1111/1467-9868.00105.

Frazier, P. A., Tix, A. P., and Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, 51(1):115. doi:10.1037/0022-0167.51.1.115.

Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401. doi:10.2307/1907330.

Fuller, W. A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28(1):5–23.

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley &Sons. doi:10.1002/9780470523551.

Geiger, C. and Kanzow, C. (2002). *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer. Springer Berlin Heidelberg. doi:10.1007/978-3-642-56004-0.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435. doi:10.1198/004017005000000661.

Gelman, A., Park, D. K., Ansolabehere, S., Price, P. N., and Minnite, L. C. (2001). Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):101–118. doi:10.1111/1467-985x.00190.

Golan, A., Judge, G., and Miller, D. (1997). *Maximum entropy econometrics: robust estimation with limited data*. Chichester (United Kingdom) John Wiley and Sons.

Gosh, M. and Meeden, G. (1997). *Bayesian methods for finite population sampling*. London: Chapman & Hall. doi:10.1007/978-1-4899-3416-1.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Guandalini, A. and Tillé, Y. (2017). Design-based estimators calibrated on estimated totals from multiple surveys. *International Statistical Review*, 85(2):250–269. doi:10.1111/insr.12160.

Hansen, M., Hurwitz, W., and Madow, W. (1953). *Sample survey methods and theory*, volume I and II. New York: Wiley. doi:10.2307/2343344.

Hanson, R. H. (1978). *The current population survey: design and methodology*, volume 40. Department of Commerce, Bureau of the Census.

Haziza, D. and Beaumont, J.-F. (2017). Construction of weights in surveys: a review. *Statistical Science*, 32(2):206–226. doi:10.1214/16-sts608.

Hidiroglou, M. A., Särndal, C.-E., and Binder, D. A. (1995). Weighting and estimation in business surveys. *Business Survey Methods*, pages 475–502. doi:10.1002/9781118150504.ch25.

Holt, D. and Smith, T. F. (1979). Post stratification. *Journal of the Royal Statistical Society. Series A (General)*, 142(1):33–46. doi:10.2307/2344652.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685. doi:10.2307/2280784.

Houbiers, M. (2004). Towards a social statistical database and unified estimates at statistics netherlands. *Journal of Official Statistics*, 20(1):55.

Huang, E. and Fuller, W. (1978). Non-negative regression estimation in sample survey data. *Proccedings Social Statitics Section, American Statistical Association*, pages 300–305.

Husain, M. (1969). Construction of regression weights for estimation in sample surveys. unpublished m.s. thesis, Iowa State University, Ames, Iowa.

Ip, E. (2001). Visualizing multiple regression. *Journal of Statistics Education*, 9(1):1–10. doi:10.1080/10691898.2001.11910646.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96. doi:10.2307/2287773.

Isaki, C. T., Tsay, J. H., and Fuller, W. A. (2004). Weighting sample data subject to independent controls. *Survey Methodology*, 30(1):35–44.

Judd, C. M. and Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5):602–619. doi:10.1177/0193841x8100500502.

Kabzinska, E. and Berger, Y. G. (2015). Aligning estimates from different surveys using empirical likelihood methods. (Unpublished Paper). University of Southampton.

Kennedy, P. (2002). More on venn diagrams for regression. *Journal of Statistics Education*, 10(1):1–10. doi:10.1080/10691898.2002.11910547.

Kennedy, P. E. (1981). The ballentine: a graphical aid for econometrics. *Australian Economic Papers*, 20(37):414–416. doi:10.1111/j.1467-8454.1981.tb00368.x.

Kennel, T. (2013). *Topics in model-assisted point and variance estimation in clustered samples*. PhD thesis, University of Maryland.

Kennel, T. and Valliant, R. (2010). Logistic generalized regression (lgreg) estimator in cluster samples. *Section on Survey Research Methods*.

Keyfitz, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association*, 52(280):503–510. doi:10.2307/2281699.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78(1):21–39. doi:10.1111/j.1751-5823.2010.00099.x.

Kish, L. (1965). *Survey sampling*. John Wiley and Sons. doi:10.2307/2283653.

Knottnerus, P. and van Duin, C. (2006). Variances in repeated weighting with an application to the Dutch labour force survey. *Journal of Official Statistics*, 22(3):565.

Kott, P. S. (2003). A practical use for instrumental-variable calibration. *Journal of Official Statistics*, 19(3):265.

Kott, P. S. (2005). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129(1-2):263–277. doi:10.1016/j.jspi.2004.06.052.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21(1):25–32.

Lehtonen, R. and Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys*. John Wiley & Sons.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24:51–56.

Lehtonen, R. and Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. In *Handbook of statistics*, volume 29, pages 219–249. Elsevier. doi:10.1016/s0169-7161(09)00231-4.

Lemaître, G. and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13:199–207.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466):546–556. doi:10.1198/016214504000000467.

Lohr, S. (2009). *Sampling: design and analysis*. Nelson Education.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010. doi:10.2307/2283327.

Luery, D. (1986). Weighting survey data under linear constraints on the weights. *In Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 325–330.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Routledge.

MacKinnon, D. P., Fairchild, A. J., and Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58(593):1–22. doi:10.1146/annurev.psych.58.110405.085542.

McConville, K. and Breidt, F. (2013). Survey design asymptotics for the model-assisted penalised spline regression estimator. *Journal of Nonparametric Statistics*, 25(3):745–763. doi:10.1080/10485252.2013.780057.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association,University of Wollongong*, 99(468):1131–1139. doi:10.1198/016214504000000601.

Mittelhammer, R. C. (2013). *Mathematical Statistics for Economics and Business*. New York: Springer. doi:10.1007/978-1-4614-5022-1.

Mohl, S. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22(2).

Montanari, G. (1987). Post-sampling efficient qr-prediction in large-sample surveys. *International Statistical Review*, 55(2):191–202. doi:10.2307/1403195.

Montanari, G. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472):1429–1442. doi:10.1198/016214505000000141.

Münnich, R. (2008). Varianzschätzung in komplexen Erhebungen. *Austrian Journal of Statistics*, 37(3&4):319–334.

Münnich, R., Burgard, J. P., Gabler, S., Ganninger, M., and Kolb, J.-P. (2012a). *Statistik und Wissenschaft: Stichprobenoptimierung und Schätzung im Zensus 2011*, volume 21. Statistisches Bundesamt.

Münnich, R., Burgard, J. P., Gabler, S., Ganninger, M., and Kolb, J.-P. (2016). Small area estimation in the German census 2011. *Statistics in Transition New Series*, 17(1):25–40. doi:10.21307/stattrans-2016-004.

Münnich, R., Burgard, J. P., and Vogt, M. (2013). Small Area-Statistik: Methoden und Anwendungen. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 6(3-4):149–191. doi:10.1007/s11943-013-0126-1.

Münnich, R., Burgard, P., and Rupp, M. (2018). A generalized calibration approach ensuring coherent estimates with small area constraints. (Unpublished Paper). Trier University.

Münnich, R., Sachs, E. W., and Wagner, M. (2012b). Calibration of estimator-weights via semismooth Newton method. *Journal of Global Optimization*, 52(3):471–485. doi:10.1007/s10898-011-9759-1.

Nangsue, N. and Berger, Y. G. (2014). Optimal regression estimator for stratified two-stage sampling. In *Contributions to Sampling Statistics*, pages 167–177. Springer.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of Indian Society of Agricultutral Statistics*, 3:169–175.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625. doi:10.1007/978-1-4612-4380-9_12.

Nieuwenbroek, N. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. *Netherlands Central Bureau of Statistics, Department of Statistical Methods*.

OECD (2007). Glossary of statistical terms. Retrieved from https://stats.oecd.org/glossary/download.asp. visited 20/09/2018.

ONS (2012). *Integrated Household Survey User Guide*. South Wales, volume 1: ihs background & methodology edition.

Park, M. and Fuller, W. A. (2005). Towards nonnegative regression weights for survey samples. *Survey Methodology*, 31(1):85–93.

Park, M. and Fuller, W. A. (2009). The mixed model for survey regression estimation. *Journal of Statistical Planning and Inference*, 139(4):1320–1331. doi:10.1016/j.jspi.2008.02.021.

Rao, J. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10(2):153.

Rao, J. (2003). *Small Area Estimation*. New York: Wiley. doi:10.1002/0471722189.

Rao, J. and Singh, A. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methodes*, American Statistical Association:57–65.

Renssen, R. H. and Nieuwenbroek, N. J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92(437):368–374. doi:10.1080/01621459.1997.10473635.

Riede, T. (2013). Weiterentwicklung des Systems der amtlichen Haushaltsstatistiken. In *Weiterentwicklung der amtlichen Haushaltsstatistiken*, pages 13–29. Riede, Thomas and Bechthold Sabine and Notburga, Ott, Scivero Verlag Berlin.

Robinson, P. and Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā: The Indian Journal of Statistics, Series B*, 45(2):240–248.

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357. doi:10.1093/ije/dyr082.

Rottach, R. A. and Hall, D. W. (2005). Using equalization constraints to find optimal calibration weights. *American Statistical Association, Section on Survey Research Methods*.

Royall, M. (1992). Robustness and optimal design under prediction models for finite populations. *Survey Methodology*, 18(2):179–185.

Royall, R. (1976). Current advances in sampling theory: implications for human observational studies. *American Journal of Epidemiology*, 104(4):463–474. doi:10.1093/oxfordjournals.aje.a112317.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2):377–387. doi:10.2307/2334846.

Särndal, C. E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67(3):639–650. doi:10.1093/biomet/67.3.639.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2):99–119.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Science & Business Media. doi:10.1007/978-1-4612-4378-6.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3):527–537. doi:10.1093/biomet/76.3.527.

Särndal, C.-E., Thomsen, I., Hoem, J. M., Lindley, D. V., Barndorff-Nielsen, O., and Dalenius, T. (1978). Design-based and model-based inference in survey sampling [with discussion and reply]. *Scandinavian Journal of Statistics*, 5(1):27–52.

Schaich, E. and Münnich, R. (2001). *Mathematische Statistik für Ökonomen*. Vahlen.

Schlittgen, R. (2008). *Einführung in die Statistik: Analyse und Modellierung von Daten*. Walter de Gruyter. doi:10.1524/9783486715910.

Seber, G. A. F. (1977). *Linear Regression Analysis*. Wiley series in probability and mathematical statistics. John Wiley & Sons. doi:10.1002/9780471722199.

Singh, A. (1996). Combining information in survey sampling by modified regression. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, volume 91, pages 120–129.

Singh, A. and Mohl, C. (1996). Understanding calibration estimators in survey sampling. *Survey methodology*, 22(2):107–115.

Snijders, T. (2011). *Multilevel analysis*. Springer. doi:10.1007/978-3-642-04898-2_387.

Snijders, T. A. and Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage.

Statistics Canada (2017). *Methodology of the Canadian Labour Force Survey*. Ottawa: Minister of Industry, fourth edition edition. Catalogue no. 71-526-X.

Steel, D. G. and Clark, R. G. (2007). Person-level and household-level regression estimation in household surveys. *Surveys Methodology*, 33(1):55–60.

Stryhn, H., De Vliegher, S., and Barkema, H. (2006). Contextual multilevel models: effects and correlations at multiple levels. In *Proceedings of the 11th International Symposium on Veterinary Epidemiology and Economics. Cairns, Australia*.

Stukel, D., Hidiroglou, M., and Särndal, C.-E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus taylor linearization. *Survey Methodology*, 22(2):117–126.

Tam, S. (1995). Optimal and robust strategies for cluster sampling. *Journal of the American Statistical Association*, 90(429):379–382. doi:10.1080/01621459.1995.10476523.

Théberge, A. (2000). Calibration and restricted weights. *Survey Methodology*, 26(1):99–108.

Thompson, S. K. (2002). *Sampling*. Wiley. doi:10.1002/9781118162934.

Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: simple random sampling. *International Statistical Review*, 66(3):303–322. doi:10.1111/j.1751-5823.1998.tb00375.x.

Väisänen, P. (2002). Estimation procedure of the Finnish time use survey 1999-2000. In *Paper to be presented at the IATUR Annual Conference*, pages 15–18.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88(421):89–96. doi:10.1080/01621459.1993.10594298.

Valliant, R. (2002). Variance estimation for the general regression estimator. *Survey methodology*, 28(1):103–108.

Valliant, R., Dever, J. A., and Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. Springer. doi:10.1007/978-1-4614-6449-5.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: a Prediction Approach*. Wiley Series in Probability and Statistics. New York: Wiley.

van den Brakel, J. (2013). Sampling and estimation techniques for household panels. Technical report, Discussion paper 2013-15, Statistics Netherlands, Heerlen.

van den Brakel, J. (2016). Register-based sampling for household panels. *Survey Methodology*, 42(1):137–159.

van den Brakel, J. and Bethlehem, J. (2008). Model-based estimation for official statistics. *Statistics Netherlands Discussion paper*.

Verma, V., Betti, G., and Ghellini, G. (2006). *Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC*. Università di Siena, Dipartimento di metodi quantitativi.

Verma, V. and Clémenceau, A. (1996). Methodology of the european community household panel. *Statistics in Transition*, 2(7):1023–1062.

von Auer, L. (2007). *Ökonometrie*. Berlin Heidelberg: Springer. doi:10.1007/978-3-642-19995-0.

Wallgren, A. and Wallgren, B. (2007). *Register-based statistics: administrative data for statistical purposes*, volume 553. John Wiley & Sons.

Wittenberg, M. (2010). An introduction to maximum entropy and minimum cross-entropy estimation using Stata. *Stata Journal*, 10(3):315–330. doi:10.1177/1536867x1001000301.

Wolter, K. (2007). *Introduction to variance estimation*. Statistics for Social and Behavioral Sciences. Springer, New York.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66(334):411–414. doi:10.2307/2283947.

Wooldridge, J. M. (2013). *Introductory econometrics: a modern approach*. South-Western College Pub.

Wright, R. L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78(384):879–884. doi:10.1080/01621459.1983.10477035.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics*, 32(1):15–26. doi:10.2307/3315996.

Wu, S., Kennedy, B., and Singh, A. C. (1997). Household-level versus person-level regression weight calibration for household surveys. *Proceedings of the Survey Methdods Section, Statistical Society of Canada*, pages 99–104.

Zhang, L.-C. (2000). Post-stratification and calibration - a synthesis. *The American Statistician*, 54(3):178–184. doi:10.1080/00031305.2000.10474542.

Zieschang, K. D. (1986). A generalized least squares weighting system for the consumer expenditure survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 64–71.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, 85(412):986–1001. doi:10.1080/01621459.1990.10474969.