ABSTRACT

Title of Dissertation: A UNIFYING PARAMETRIC FRAMEWORK FOR ESTIMATING FINITE POPULATION TOTALS FROM COMPLEX SAMPLES

Ismael Flores Cervantes, Doctor of Philosophy, 2019

Dissertation directed by: Dr. J. Michael Brick, Research Professor
Dr. Frauke Kreuter, Professor
Joint Program in Survey Methodology, University of Maryland

We propose a unifying framework for improving the efficiency of design-based estimators of finite population characteristics in the presence of full response. We call the framework a *Parametric* (PA) approach. The PA framework, an extension of the model-assisted theory, uses an algorithmic approach driven by the observed data. The algorithm identifies the relevant subset of auxiliary variables related to the outcome, and the known population totals of these variables are used to compute the PA estimator. We apply the PA framework to three important estimation problems: the identification of the functional form of a design-based estimator based on the observed data; the identification working or assisting model; and the development of the methodology for creating new design-based estimators. The PA estimators are theoretically justified and evaluated by simulations. This dissertation is limited to

single-stage sample designs with full response, but the framework can be extended to other sample designs and for estimation with nonresponse.

A UNIFYING PARAMETRIC FRAMEWORK FOR ESTIMATING FINITE POPULATION TOTALS FROM COMPLEX SAMPLES

by

Ismael Flores Cervantes

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor in Philosophy
2019

Advisory Committee:
Dr. J. Michael Brick, Co-Chair
Dr. Frauke Kreuter, Co-Chair
Dr. S. Lynne Stokes
Dr. Richard Valliant
Dr. Paul Smith
Dr. Partha Lahiri

Preface

Survey sampling, with its own theory and methodology, has been considered as a small niche within standard statistics. This situation has produced a disconnect between theory and practice. For example, nonresponse is one of the most important challenges facing survey sampling theory; however, most textbooks dedicate only a few pages to this problem. As noted by Tillé (2006), a concept as common as simple random sampling is often not defined, although it can be described mathematically as a discrete random vector with a probability density mass and a characteristic function.

In this dissertation, we call for a change of perspective in the current approach to estimation in survey sampling. We extend Tillé's idea and postulate that sample designs are uniquely defined as a multivariate discrete random variable with an expected value and a variance-covariance matrix with specific properties that determine the type of design. The observed sample is also a multivariate discrete distribution with a probability mass function that inherits the properties from the random vector that describes the sample selection. Furthermore, all estimators are functions of these random variables. Since there are no differences between a sample design and design-based estimators and other random variables and functions of random variables, we can use standard statistical analysis for studying design-based estimators. This approach justifies the use of tools from standard statistics and other fields such as engineering and physics. The introduction of matrix notation and

matrix operations provides new insights into the performance of estimators without the use of simulations.

As shown in this research, the proposed methodology, called *Parametric* (PA) Approach, has been useful for the design of algorithmic estimators that address the problem of working model building and variable selection for calibration. The algorithm was engineered based on the observations of the mathematical relationship between the outcome variable and the probability of inclusion using orthogonal components, a tool commonly used in other fields. Under this approach, we have a better understanding of when estimators are efficient or when they underperform. These ideas also provide a methodology to develop new design-based estimators from any model that is capable of reproducing the classical design-based estimators. Using the same tools, we revisit the survey sampling asymptotic theory and provide a more intuitive way to study the large sample properties of estimators. We also revisited some unreproducible results reported in the literature.

The main consequence of this change in perspective is the rethinking of concepts such as the role of models within the design-based paradigm while questioning engrained concepts in the current theory. However, developing a new unifying framework is not the goal of this endeavor. The main goal is to provide tools for addressing the current problems facing the field.

Dedication

I dedicate this work to my mother Socorro Cervantes Garcia who always supported me and helped me type this dissertation. To the memory of my father David Flores Sierra who would have been very proud of this accomplishment.

I thank both my parents for supporting me in reaching my goals and their encouragement along the way.

To my brothers David, Emilio, Ruben, and Edgar, who are also my best friends. Thank you for always being there for me.

To Russell who patiently stood by me despite the chaos of books, computers, and papers all over the floor. Thank you for your help in the last step of this work.

To my teachers la Madre Esperanza, la Maestra Charito, la Maestra Celia, el Profesor Angel, el Capi Oscar, el Capi Fermin, el Profe Rene (el Pazuzu), el Profe Varela, el Doc Arroyo, Don Pedro Palou, Billy, Juan Jose Rosales (el Tlacuache), la Seño Elvira, la Seño Josefina Olvera, la Seño Yolanda, la Seño Raquel Olga, Dr. Baez, and Mrs. Dorothy Teff. Thank you for your guidance.

To Lynne Stokes who opened the door to the survey sampling world. Thank you for your guidance and friendship. I wish I had done this sooner.

To Mike Brick for his guidance, encouragement, and friendship, not only for this work, but also for my career.

To Roger Tourangeau who convinced me to continue and reach this goal. Thank you for your guidance and friendship.

To Dr. Larry, thank you for your support and encouragement. I don't think I could have done it without your help.

I had a blast doing this work. It was really fun!

Acknowledgments

I want to express my deepest appreciation and gratitude to my thesis advisor, Dr. J. Michael Brick. This research would not have been possible without his guidance, expertise, dedication, mentoring, patience, and support in every step of this work.

I want to acknowledge my committee members, Dr. Frauke Kreuter, Dr. Lynne Stokes, Dr. Richard Valliant, Dr. Partha Lahiri, and Dr. Paul Smith.

I also want to acknowledge the insightful conversations with Dr. Graham Kalton and Dr. Roger Tourangeau who helped me see different points of view.

Finally, I appreciate the support of my family, friends, and coworkers (unnamed here) who have cheered me along the way. I am very grateful for all my colleagues and friends.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1 The Parametric Approach to Survey Sampling Estimation

1.1 Introduction

This dissertation extends the model-assisted theory for estimating enumerative finite population characteristics such as totals and means from complex survey data in the presence of full response. In the model-assisted approach, the working model for the outcome variable guides the form of the estimator, and the inferences are design-based (Särndal, Swensson, & Wretman, 1992). This approach allows for incorporating auxiliary information to improve the efficiency of the estimators. Although the working model does not need to be true for design-consistency, the gain in efficiency depends on how well the model fits the observed data.

We propose a new framework for developing design-based estimators of finite population characteristics called a *Parametric* (PA) approach in the presence of full response. The PA framework is a data-driven methodology for (1) developing the working model (i.e., choosing the auxiliary variables and functional form of the model) given the realized sample, and (2) incorporating the auxiliary variable population totals directly into the model. Unlike most design-based estimators, the PA estimator is not a single estimator, but a class of estimators called algorithmic estimators that result from applying an unambiguous set of steps or procedures to the observed sample. The PA framework is similar to, and motivated in part by, the data-

driven methods from statistical learning theory (Hastie, Tibshirani, & Friedman, 2009).

As an algorithmic-based methodology, the PA framework has these key steps.

1. Postulate a collection of well-defined parametric working models based on the available auxiliary variables. Two models are considered. The first is the standard model of the outcome variable(s). The second is a model of the probabilities of inclusion, even though these may be known. This second modeling activity differs from the model-assisted paradigm. The rationale for modeling inclusion probabilities is three-fold. First, the estimated probabilities may produce more efficient estimators than those using the known probabilities (Lumley, Shaw, & Dai, 2011). Second, the modeled probabilities of selection can stabilize estimators such as the Horvitz-Thompson (HT) estimator (Horvitz & Thompson, 1952) in some designs (Rao, 1966); for example, estimators with poststratified weight to the total population size. Third, this modeling step is essential when uncontrolled nonresponse is present, although this topic is not addressed here.

2. Evaluate the goodness of fit for both models and then identify the common variables that explain both the outcome variable and the inclusion probabilities in both models.

3. Refit a model of the inclusion probabilities using only the common variables that explain the outcome variable and inclusion probabilities. Using this model,

predict the fitted mean of the inclusion probabilities and adjust the original sample design weights.

4. Using the adjusted weights from the previous step, evaluate the goodness of fit of the models of the outcome variable to identify the auxiliary or predictor variables of the model that give the best fit.

5. Fit a model for the outcome variable using the predictors identified in the previous step using the original sampling weights and then adjust the regression coefficients of the parameters of this model using population totals of the selected auxiliary variables.

6. Construct the PA estimator as the weighted sum of the adjusted pseudo-maximum-likelihood (PML) estimates of the mean of the selected working model and estimate its variance.

Although the PA estimators are solutions of the likelihood of parametric models, we show that they are design-consistent irrespective of the fit of the working model, and the inference depends only on how the sample is drawn. Since the algorithm measures the goodness of fit of the models, the resulting PA estimator is likely to be one of the most efficient estimators among those from the evaluated working models. Because the algorithm defines the PA estimator, the asymptotic properties such as design consistency under suitable regularity conditions are given using the generic form of the PA estimator.

The PA framework uncovers interesting relationships between some PA estimators under specific models and well-known, design-based estimators. Most classical design-based estimators are shown to be weighted sums of adjusted PML estimates of parameters of the assumed working model. This relationship between the estimators and their parametric working models justifies the use of standard statistical modeling techniques within the design-based context in the PA framework.

The PA framework is applied in this paper to address three estimation problems reported in the literature. The first problem is the identification of the functional form of a design-based estimator based on the observed data. The second problem is the identification of the variables that should be used in calibration. This problem is also known as working or assisting model development. The third problem is the methodology to develop new design-based estimators. The PA algorithm provides a recipe for deriving new estimators. Since the PA framework provides a guide to "engineer" new estimators, we propose an alternative estimator for Poisson sampling designs, and two new classes of estimators called algebraic PA estimators and non-linear PA algorithmic estimators. All these PA estimators only require the auxiliary variable population totals. We evaluate and compare the PA estimator to alternative estimators described in the literature using simulations by varying factors such as sample design, working model misspecification, and sample size.

1.2 Background and the Need for Change

The survey sampling literature describes numerous estimators for finite population characteristics that rely on auxiliary information to improve the efficiency of the HT estimator. These estimators are constructed by assuming that the underlying working model is known and correctly specified. Frequently, the estimators are evaluated under optimal conditions (e.g., the working model is correctly specified) through simplistic simulations. Little guidance is available for identifying the auxiliary variables in the model, nor are diagnostics given to determine if the underlying assumptions hold. As a result, it is difficult to assess the efficiency of the proposed estimators in practical situations. For example, calibration estimators have been shown to be efficient compared to estimators based on PML estimators (Kott, 2006; Kim & Riddles, 2012). However, in practice, some calibration estimators may not be feasible, the auxiliary variables may have low predictive power, or the auxiliary variables may have to be selected from a large pool of variables without any guidance. It is unclear if the calibration estimators would be better in these situations.

Most current research searches for the functional form of the best estimator in a particular situation, often leading to a single functional form or expression of the estimator. However, this approach does not recognize that no single estimator works well for all conditions and sampling strategies (Rao, 2008). Another issue is that survey statisticians do not have a predetermined set of auxiliary variables for their working models and must rely on some form of data dredging to identify these

variables. Addressing these issues requires a new approach that can adapt the estimation process to what the sample or observed data reveals about the population.

The PA framework does not assume that the working model is known; instead, it focuses on the methodology for model development or model building based on the observed data. The PA estimators are the result of an algorithmic process where a single form of the estimator may not even exist under repeated sampling. In this regard, the PA methodology is similar to the Targeted Maximum Likelihood Estimation (TMLE) for observational studies (van der Laan & Rose, 2011), and the Double Machine Learning (DML) for treatment and causal parameters (Chernozhukov et al., 2017). The PA approach, the TMLE, and DML methodologies only target model parameters related to the outcome. The PA approach differs from the TMLE and DML because it uses these parameter estimates to produce design-based estimators of finite population characteristics.

1.3 Example of an Algorithmic PA Estimator

The PA framework for estimation with full response provides tools to determine the best functional form of an estimator from the single realization of the sample and the set of auxiliary variables that should be used in the estimator.[1] To illustrate the use of

[1] The best functional form of an estimator and the set of auxiliary variables are related because the auxiliary variables determine the form of the estimator. See Section 3.1.

the PA methodology, we use the example from Section 14.3.2 in Valliant, Dever, & Kreuter (2013) denoted as VDK to compute two different algorithmic PA estimators.

EXAMPLE 1.1. VDK discusses the selection of covariates as control totals for generalized linear regression estimators (GREG) (see Cassel, Särndal, & Wretman, 1977). VDK illustrates the differences in efficiency of GREG estimators using different sets of auxiliary variables by computing two estimators using the 1998 Survey of Mental Health Organizations data set `smho.N874` from the R package `PracTools` (Valliant, Dever, & Kreuter, 2018). The renamed variables with renumbered levels and their description from the file `smho.N874` used in this example are listed in Table 1.1.

In the VDK example, the population consists of $N = 725$ hospitals[2] and a systematic sample of $n = 80$ hospitals is selected with a probability proportional to size (PPS) from the frame randomly ordered before sample selection. The measure of size (MOS) of a hospital is $m_k = \sqrt{5 + \mathbf{1}_{\{x_{2k}>5\}}(x_{2k} - 5)}$ for $k \in U$ where $\mathbf{1}_{\{x_{2k}>5\}}$ is the indicator function for $x_{2k} > 5$ where $\mathbf{1}_{\{x_{2k}>5\}} = 1$ if $x_{2k} > 5$ or $\mathbf{1}_{\{x_{2k}>5\}} = 0$ if $x_{2k} \le 5$, and x_{2k} is the number of inpatient beds in hospital k for $k \in U$. The inclusion

[2] The original frame is the file `smho.N874` with 874 hospitals but 149 records coded as `hosp.type`=4 for outpatient and partial cases hospitals are removed before the analysis. The variable x_1 contains the renumbered levels of `hosp.type` and $\mathbf{x}_1$ is the vector of dummy variables for each hospital type as indicated in Table 1.1.

Table 1.1 Variables in the frame from the 1988 Survey of Mental Health Organizations

Variable	Type	Description	Levels/values
y_1	Dependent /continuous	Hospital total expenditures in 1998	99,000 to 197,210,630
y_2	Dependent /binary	Indicator for whether the hospital received financing from the state mental health agency in 1998	$y_2 = 1$: Hospital received financing $y_2 = 0$: Hospital did not receive financing
$\mathbf{x}_1 = (x_{11}, x_{12}, x_{13}, x_{14})$	Auxiliary /categorical	Hospital type	$x_{1,1} = 1$: Psychiatric, $x_{1,1} = 0$: Otherwise $x_{1,2} = 1$: Residential/ veterans, $x_{1,2} = 0$: Otherwise $x_{1,3} = 1$: General, $x_{1,3} = 0$: Otherwise $x_{1,4} = 1$: Multiservice/ substance abuse $x_{1,4} = 1$: Otherwise
x_2	Auxiliary /discrete (assumed continuous)	Total inpatient beds	0 to 1,357
x_3	Auxiliary /discrete (assumed continuous)	Unduplicated client/ patient seen during the year	0 to 28,993
x_4	Auxiliary /discrete (assumed continuous)	End of year count of patients on the roll	0 to 14,239

probability is $\pi_k = n\frac{m_k}{\sum_{k\in U} m_k}$ where n is the sample size. We use the same random seed for the sample selection to reproduce the results from VDK for the comparison with the algorithmic PA estimators.

The first VDK estimator is $\hat{Y}_{VDK,1}$, the estimator of the total expenditures in 1998 for all hospitals in the frame, Y_1, based on the variable y_1, which is the individual hospital expenditures. The second estimator is $\hat{\bar{Y}}_{VDK,2}$, the estimator of the proportion of hospitals that received financing from the state mental health agency in 1998, $\bar{Y}_2 = \frac{Y_2}{N}$, based on the variable y_2, which is the indicator of whether or not the hospital received financing from the state agency. The population totals of the auxiliary variables of the estimators $\hat{Y}_{VDK,1}$ and $\hat{\bar{Y}}_{VDK,2}$ are $\left(N, \mathbf{X}_1, X_3, X_4, \mathbf{X}_5\right)$ and $\left(N, \mathbf{X}_1, X_2, X_3, X_4\right)$, respectively, where $\mathbf{X}_5 = \mathbf{X}_1 * X_2$ represents the population totals of the interaction between the variables $\mathbf{x}_1$ and x_2; that is, the total number of beds by hospital type shown in Table 1.2. (See Section 1.5.2 for notation of models and variables).

VDK selected the auxiliary variables for $\hat{Y}_{VDK,1}$ using the results of an analysis of the dependent variable y_1 based on the full population. After fitting a generalized linear model (GLM) to the outcome y_1 and examining the slope of the variable x_2 (number of beds) by $\mathbf{x}_1$ (hospital type), they decided to include these variables as main effects

in the working model of $\hat{Y}_{VDK,1}$. Their population analysis for y_2 showed different slopes by hospital type so $\mathbf{x}_1$ (hospital type) was selected as the main effects and the interaction terms between $\mathbf{x}_1$ and x_2 were excluded from the working model of $\hat{\bar{Y}}_{VDK,2}$. These analyses are not possible in practice since the dependent variables are only observable for the sampled cases after sample selection.

Table 1.2 Auxiliary variable for the number of inpatient beds by hospital type, $\mathbf{x}_5 = \mathbf{x}_1 * x_2$ where $\mathbf{x}_1$ is hospital type and x_2 is the total inpatient beds in the hospital in the 1988 Survey of Mental Health Organizations data

Variable	Levels/values
$\mathbf{x}_{k5} = \mathbf{x}_{k1} * x_{k2}$ $= (x_{k51}, x_{k52}, x_{k53}, x_{k54})$	$x_{k51} = x_2$: If hospital k is psychiatric, $x_{k51} = 0$: Otherwise. $x_{k52} = x_2$: If hospital k is residential/ veterans, $x_{k52} = 0$: Otherwise. $x_{k53} = x_3$: If hospital k is general, $x_{k52} = 0$: Otherwise. $x_{k54} = x_4$: If hospital k is multiservice/ substance abuse, $x_{k52} = 0$: Otherwise.

The goal of the PA algorithm is to identify the relevant variables that explain the outcome variable from the observed sample considering the sample selection. After these variables are identified, the algorithm incorporates the population totals of these variables into the pseudo-log-likelihood (PLL) of the data for an assumed working model with these variables. This information is currently ignored in the regular PML

approach (Binder & Roberts, 2009). Then the PA estimator is derived as the sum of the expanded adjusted fitted means of the working model.

We describe how to compute two separate algorithmic estimates of the total Y_1, $\hat{Y}_{pa,1}$, and the proportion $\bar{Y}_2$, $\hat{\bar{Y}}_{pa,2}$, using the PA approach. As in the VDK example, we expect to use different sets auxiliary variables in the PA working models of Y_1 and $\bar{Y}_2$. The PA estimators $\hat{Y}_{pa,1}$ and $\hat{\bar{Y}}_{pa,2}$ are derived following the steps of Algorithm 1.1 on page 61. The algorithm consists of 10 steps classified into four separate groups with specific goals:

A. Identification of the best-fit Maximum Likelihood/ Pseudo-Maximum Likelihood working models of the outcome variable and probabilities of inclusion (Steps 1 to 4).

The PA algorithm starts by fitting separate models for the sample membership indicator S_k in Steps 1 and 2, and the outcome variable y_k in Steps 3 and 4 to identify a working model with the auxiliary variables that are predictors of both the probability of inclusion π and the outcome variable y.

STEP 1. Propose the collection of working models $\boldsymbol{\mathcal{M}}_{\pi}$ for the sample membership indicator S_k for $k \in U$.

In the first step, we define the distribution function of the working model for S_k. In this example, we assume that the population is available (see Section 1.6 for alternatives for modeling S_k when only the sample is available). Let $S_k \in \{0,1\}$ be a discrete random variable for the sample membership indicator and s_k be the realization of S_k (e.g., $S_k = s_k$) that takes the value of one if the unit k is selected in the sample or zero if the unit k is not selected for $k \in U$. Let $\mathbf{S} = [S_k] \in (0,1)^{N \times 1}$ be the discrete random vector with the sample membership indicator S_k for all the elements in the population. We assume that the observed sample (e.g., all cases with $s_k = 1$) is a realization of S_k for $k \in U$, which is assumed to follow a Bernoulli distribution $S_k \overset{iid}{\sim} \mathcal{B}e(\pi_k)$ where $\pi_k = \text{logit}^{-1}(\mathbf{x}_k \boldsymbol{\beta})$, $\mathbf{x}_k \in \mathbb{R}^{1 \times P}$, $\mathbf{x} = (x_1, ..., x_P)$ is the vector of auxiliary variables associated with the element $k \in U$, $\boldsymbol{\beta} \in \mathbb{R}^{P \times 1}$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_P)^{\mathrm{T}}$ is the vector of the regression coefficients, and T is the transpose operator.

Let $\mathcal{M}_{\pi}$ be the true model for $\mathbf{S}$ and $\boldsymbol{\mathcal{M}}_{\pi}$ the set or collection of working models for $\mathbf{S}$ generated by the linear combinations of the auxiliary variables $\mathbf{x} = (x_1, ..., x_P)$ and any values of $\boldsymbol{\beta} \in \mathbb{R}^{1 \times P}$ (see Definitions 1.1 and 1.3). In this example, the vector of auxiliary variables is $\mathbf{x} = (1, \mathbf{x}_1, x_2, x_3, x_4, \mathbf{x}_5)$. The population totals for the models

in $\mathcal{M}_\pi$ are the combination of the totals $\mathbf{X}=\left(N,\mathbf{X}_1,X_2,X_3.X_4,\mathbf{X}_5\right)$. There are an infinite number of models in $\mathcal{M}_\pi$ and none of the models in $\mathcal{M}_\pi$ is correctly specified since the true model of $\mathbf{S}$ is a nonlinear function of x_{k1}. However, the algorithm does not require the correct working model of $\mathbf{S}$ because the model is only used to identify the relevant auxiliary variables that explain the sample selection. Since the models in $\mathcal{M}_\pi$ are defined at the population level, the parameters of these models are estimated using Maximum Likelihood (ML) where the sampling weights do not play any role in the estimation (Casella & Berger, 2002).

The key outcome of Step 1 is $\mathcal{M}_\pi$, the collection of working models for the sample membership indicator S_k.

STEP 2. Identify the ML model $\widehat{\mathcal{M}}_\pi \in \mathcal{M}_\pi$ for $\mathbf{S}$ that minimizes the loss function $\mathrm{L}(S)$.

The expression for the log likelihood (LL) of the models for $\mathbf{S}=\mathbf{s}$ in $\mathcal{M}_\pi$ fitted to the complete population is

$$\log \mathcal{L}(\boldsymbol{\beta};\mathbf{S},\mathbf{x})=\sum_{k\in U}\left(S_k\mathbf{x}_k\boldsymbol{\beta}-\log\left(1+\exp\left(\mathbf{x}_k\boldsymbol{\beta}\right)\right)\right), \qquad (1.1)$$

where $\boldsymbol{\beta}=\left(\beta_1,...,\beta_P\right)^{\mathrm{T}}$ are the regression coefficients for the auxiliary variables $\mathbf{x}_k=\left(x_{k1},...,x_{kP}\right)$. Let $\widehat{\mathcal{M}}_\pi$ be the set of all ML models in $\mathcal{M}_\pi$; then the maximum

likelihood estimates (MLE) of the regression coefficients, $\hat{\boldsymbol{\beta}}_{mle} \in \widehat{\boldsymbol{\mathcal{M}}}_{\pi}$ are the solutions to

$$\hat{\boldsymbol{\beta}}_{mle} = \underset{\boldsymbol{\beta} \in \boldsymbol{\mathcal{M}}_{\pi}}{\arg\max}\ \log \mathcal{L}(\boldsymbol{\beta}). \tag{1.2}$$

Let $\widehat{\mathcal{M}}_{\pi} \in \widehat{\boldsymbol{\mathcal{M}}}_{\pi}$ be the ML model for $\mathbf{S}$ among the models in $\widehat{\boldsymbol{\mathcal{M}}}_{\pi}$ with the lowest value of the loss function $\mathrm{L}(S)$. In the PA algorithm, we do not fit all ML models $\widehat{\boldsymbol{\mathcal{M}}}_{\pi} \subset \boldsymbol{\mathcal{M}}_{\pi}$ to identify the model $\widehat{\mathcal{M}}_{\pi}$; instead, we use a forward stepwise variable selection where L_{π} is the AIC, the Akaike information criterion (Akaike, 1981) to generate and fit a subset of ML models $\widehat{\boldsymbol{\mathcal{M}}}_{\pi}$ from $\boldsymbol{\mathcal{M}}_{\pi}$ (see details of the variable selection and the AIC in Section 1.7). In this example, the same ML model $\widehat{\mathcal{M}}_{\pi}$ is fitted for y_1 and y_2 since $\widehat{\mathcal{M}}_{\pi}$ does not depend on the dependent variable. The ML working model for $\mathbf{S}$ with the best fit is $\widehat{\mathcal{M}}_{\pi} = (1, x_{11}, x_{12}, x_4)$ with a loss value of $\mathrm{L}(S) = -481.46$.

The key outcome of Step 2 is $\widehat{\mathcal{M}}_{\pi}$, the ML model with the specification of the auxiliary variables of the best-fit working model of the sample membership indicator S_k for $k \in U$.

STEP 3. Propose the collection of working models $\boldsymbol{\mathcal{M}}_y$ for the outcome y.

Similar to Step 1, we first assume distribution functions for the models for the outcomes variables y_1 and y_2.

Let $\mathcal{M}_y$ be the true model for y and $\boldsymbol{\mathcal{M}}_y$ be the set or collection of working models for y where $y_k \mid \mathbf{x}_k \overset{iid}{\sim} \mathcal{N}\left(\mathbf{x}_k\boldsymbol{\beta},\sigma^2\right)$. The models in $\boldsymbol{\mathcal{M}}_y$ are generated by the linear combinations of the auxiliary variables $\mathbf{x}=\left(x_1,...,x_P\right)$ and any values of $\boldsymbol{\beta}\in\mathbb{R}^{1\times P}$. We also use this collection of models for y_2 even though they are misspecified because y_2 is a binary variable. For the PA estimators, we define the collection of models $\boldsymbol{\mathcal{M}}_{y_1}$ and $\boldsymbol{\mathcal{M}}_{y_2}$ for y_1 and y_2 using the same set of auxiliary variables $\left(1,\mathbf{x}_1,x_2,x_3,x_4,\mathbf{x}_5\right)$ and population totals $\left(N,\mathbf{X}_1,X_2,X_3.X_4,\mathbf{X}_5\right)$ in $\boldsymbol{\mathcal{M}}_\pi$ from Step 1. The models in $\boldsymbol{\mathcal{M}}_{y_1}$ and $\boldsymbol{\mathcal{M}}_{y_2}$ include those for $\hat{Y}_{1,VDK}$ and $\hat{\bar{Y}}_{2,VDK}$, in addition to the Hájek (HJ) estimator, among others.

The key outcome of Step 3 is $\boldsymbol{\mathcal{M}}_y$, the collection of models of the outcome(s).

STEP 4. Identify the PML model $\widehat{\mathcal{M}}_y \in \widehat{\boldsymbol{\mathcal{M}}}_y$ for y that minimizes the loss function L_y using the sampling weights $d_k = \frac{1}{\pi_k}$.

The expression of the PLL of the models of y_k, in $\boldsymbol{\mathcal{M}}_y$ fitted to the observed sample is

$$\log \mathcal{L}(\boldsymbol{\beta}, \sigma; \mathbf{S}, \mathbf{d}, \mathbf{x} \mid \mathcal{F}) = -\sum_{k \in U} S_k d_k \left(\log(\sigma) + \frac{\log(2\pi)}{2} + \frac{1}{2\sigma^2}(y_k - \mathbf{x}_k \boldsymbol{\beta})^2 \right), \qquad (1.3)$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_P)^{\mathrm{T}} \in \mathbb{R}^{P \times 1}$ are the regression coefficients for the auxiliary variables $\mathbf{x}_k = (x_{k1}, ..., x_{kP})$ and $d_k = \pi_k^{-1}$ are the sampling weights for $k \in U$. Let $\widehat{\boldsymbol{\mathcal{M}}}_y$ be the collection of all PML models in $\boldsymbol{\mathcal{M}}_y$ where the PMLE of the regression coefficients $\hat{\boldsymbol{\beta}}_{pmle} \in \widehat{\boldsymbol{\mathcal{M}}}_y$ are the solutions to

$$\hat{\boldsymbol{\beta}}_{pmle} = \underset{\boldsymbol{\beta} \in \boldsymbol{\mathcal{M}}_y}{\arg\max} \log \mathcal{L}(\boldsymbol{\beta} \mid \mathcal{F}). \qquad (1.4)$$

Let $\widehat{\mathcal{M}}_y \in \widehat{\boldsymbol{\mathcal{M}}}_y$ be the PML model for y among the models in $\widehat{\boldsymbol{\mathcal{M}}}_y$ with the lowest value of the loss function $\mathrm{L}(y)$. As in Step 1, we do not fit all PML models in $\widehat{\boldsymbol{\mathcal{M}}}_y$ to identify $\widehat{\mathcal{M}}_y$. We use a forward stepwise variable selection based on the *dAIC*, a sample-based estimator of the AIC, to generate and fit a subset of the PML models $\widehat{\boldsymbol{\mathcal{M}}}_y$ from $\boldsymbol{\mathcal{M}}_y$ (see details of *dAIC* in Section 1.7 and Section A.4 in Appendix A).

In this example, the PML working model for y_1 with the best fit is $\widehat{\mathcal{M}}_y(y_1) = (1, x_{11}, x_2, x_3, x_{52})$ with a loss value of $\mathrm{L}(y_1) = -2{,}843.2$. The ML working model of y_1 includes the variables x_2 (number of hospital beds), x_3 (unduplicated number of client/ patients seen during the year), the indicator x_{11} (indicator for psychiatric hospitals) and x_{52} (number of beds in residential/veterans hospitals, see Table 1.2). The PML working model for y_2 with the best fit is $\widehat{\mathcal{M}}_y(y_2) = (1, x_{12}, x_{15}, x_4, x_{52})$ with a loss value of $\mathrm{L}(y_2) = -58.40$. The working model of y_2 includes the indicators x_{12} and x_{14} (indicators for residential/veterans and multiservice/substance abuse hospitals), the variable x_4 (end of year count of patients on the hospital roll), and the variable x_{52} (number of beds in residential/veterans hospitals). The PML working model for y_2 is reasonable since substance abuse hospitals and large residential/veteran hospitals (measured by the number of beds) tend to receive funding from the state agency.

The key outcome of Step 4 is $\widehat{\mathcal{M}}_y$ with the specification of the variables of the working model of the outcome(s) with the best fit.

B. Targeting of relevant variables for y and S (Steps 5, 6, and 7)

The second group of steps of the PA algorithm (Steps 5, 6, and 7) identifies the explanatory auxiliary variables for both the outcome and the sample membership indicators. This step is done by examining the auxiliary variables in the working

models $\widehat{\mathcal{M}}_y$ and $\widehat{\mathcal{M}}_\pi$. Once the common auxiliary variables in both models are identified, a new collection of working models with these variables, $\boldsymbol{\mathcal{M}}_{\pi,y}$ for **S** is proposed. The best fit ML model working $\widehat{\mathcal{M}}_{\pi y} \in \boldsymbol{\mathcal{M}}_{\pi,y}$ is used to produce estimates of π_k, $\hat{\pi}_k$, that are used to produce estimates of the sampling weights as $\hat{d}_k = \hat{\pi}_k^{-1}$. The estimated sampling weights $\hat{d}_k$ are used to adjust the original sampling weights d_k to produce the adjusted weights $\hat{w}_k$. The adjusted weights, $\hat{w}_k$ ensure that the predictors of both the outcome variable and inclusion probabilities are retained in the models produced in the subsequent steps of the algorithm.

STEP 5. Identify the set of models $\boldsymbol{\mathcal{M}}_{\pi,y}$ for **S** using the auxiliary variables that explain both y and **S** as $\boldsymbol{\mathcal{M}}_{\pi,y} = \widehat{\mathcal{M}}_\pi \bigcap \widehat{\mathcal{M}}_y$.

Let $\boldsymbol{\mathcal{M}}_{\pi,y}$ be the set of models generated by common auxiliary variables that explain both y and **S**. The common auxiliary variables are the variables that appear in both $\widehat{\mathcal{M}}_\pi$ and $\widehat{\mathcal{M}}_y$ models from steps 2 and 4. In the case of simple random sampling (SRS), the common variable may be the intercept term, $\boldsymbol{\mathcal{M}}_{\pi,y} = 1$.

In this example, the models in $\boldsymbol{\mathcal{M}}_{\pi,y}$ for y_1 are

$$\boldsymbol{\mathcal{M}}_{\pi,y}(y_1) = (1, x_{11}, x_{12}, x_4) \bigcap (1, x_{11}, x_2, x_3, x_{52}) = (1, x_{11}).$$

Note that all working models for **S** in $\mathcal{M}_{\pi,y}(y_1)$ have a distribution $\mathcal{B}e(\pi_k)$ with $\pi_k = \frac{\exp(\beta_0 + \beta_{11}x_{k11})}{1+\exp(\beta_0 + \beta_{11}x_{k11})}$ where $\beta_0 \neq 0$ and $\beta_{11} \neq 0$. The relevant predictors for both π and y_1 are the auxiliary variables $(1, x_{11})$.

The models in $\mathcal{M}_{\pi,y}$ for y_2 are

$$\mathcal{M}_{\pi,y}(y_2) = (1, x_{11}, x_{12}, x_4) \bigcap (1, x_{12}, x_{15}, x_4, x_{52}) = (1, x_{12}, x_4),$$

where all working models for **S** in $\mathcal{M}_{\pi,y}(y_2)$ have a distribution $\mathcal{B}e(\pi_k)$ with $\pi_k = \frac{\exp(\beta_0 + \beta_{12}x_{k12} + \beta_4 x_{k4})}{1+\exp(\beta_0 + \beta_{12}x_{k12} + \beta_4 x_{k4})}$ where $\beta_0 \neq 0$, $\beta_{12} \neq 0$, and $\beta_4 \neq 0$. The relevant predictors for both **S** and y_2 are the auxiliary variables $(1, x_{12}, x_4)$. Note that relevant predictors for **S** and y_1 are not the same as the relevant predictors for **S** and y_2.

The key outcome of Step 5 is $\mathcal{M}_{\pi,y}$, the 'reduced' set of working models with the specification of the auxiliary variables that explain both the sample membership indicators and the outcome(s).

STEP 6. Fit the ML working $\widehat{\mathcal{M}}_{\pi y}$ for **S** using the auxiliary variables from the collection of models $\mathcal{M}_{\pi y}$ identified in Step 5. Using the model $\widehat{\mathcal{M}}_{\pi y}$, compute the fitted probability of selection $\hat{\pi}_k$ to produce the estimated weights $\hat{d}_k = \frac{1}{\hat{\pi}_k}$ for the

sampled units. Use these estimated weights $\hat{d}_k$ to adjust the sampling weights as $\hat{w}_k = d_k \hat{d}_k \frac{\sum_{k \in U} d_k}{\sum_{k \in U} d_k \hat{d}_k}$. The adjusted weight $\hat{w}_k$ is the expanded estimated weight $d_k \hat{d}_k$ poststratified to the total $\sum_{k \in U} d_k$. In the case of SRS, the estimated probability of selection is $\hat{\pi}_k = c$ where c is a constant, then the adjusted weight $\hat{w}_k = d_k$ which is the design weight without adjustment. In other words, for noninformative designs with respect to y and $\mathbf{S}$, there is no need to follow steps 1 to 6 of the algorithm.

We implement this step in the same way as in Step 2 but considering only the models in $\mathcal{M}_{\pi,y}$. This step is important for informative designs where the auxiliary variables used to estimate π are at the same time predictors of the outcome. The function of the adjusted weights is to ensure that the variables that explain both y and $\mathbf{S}$ are retained in the model in the following steps of the algorithm.

The key outcome of Step 6 is the adjusted sampling weight, $\hat{w}_k$.

STEP 7. Identify the PML model $\widehat{\mathcal{M}}^*_{y,\hat{w}}$ for y that minimizes the loss function $\mathrm{L}(y)$ among models in $\mathcal{M}_y$ using the adjusted weights $\hat{w}_k$ computed in Step 6.

We repeat the same procedure from Step 4 but using the adjusted weight $\hat{w}_k$ when fitting the models in $\mathcal{M}_{y,\hat{w}}$. The expressions of the PLL and $\hat{\boldsymbol{\beta}}_{pmle} \in \widehat{\mathcal{M}}^*_{y,\hat{w}}$ are given

in (1.3) and (1.4) after replacing $\mathcal{M}_y$ by $\widehat{\mathcal{M}}^*_{y,\hat{w}}$ and **d** by $\hat{\mathbf{w}}$, respectively. In this example, the PML model $\widehat{\mathcal{M}}^*_{y,\hat{w}}$ for y_1 is $\widehat{\mathcal{M}}^*_{y,\hat{w}}(y_1) = (1, x_2, x_3, x_{51})$ with a loss value of $\mathrm{L}(y_1) = -$ 2,812.4. The PML model $\widehat{\mathcal{M}}^*_{y,\hat{w}}$ for y_2 is $\widehat{\mathcal{M}}^*_{y,\hat{w}}(y_2) = (1, x_{12}, x_{14}, x_4, x_{51})$ with a loss value of $\mathrm{L}(y_2) = -$ 58.4.

The key outcome of this step is the model $\widehat{\mathcal{M}}^*_{y,\hat{w}}$ with the specification of the variables of the working model with the best fit of both the sample membership indicator and the outcome variables(s) using the weight $\hat{w}_k$.

C. Creation of the PA estimator and inference (Steps 8 to 10)

In Step 8, the final PLL model for y, $\widehat{\mathcal{M}}^*_y$, is fitted using the sampling weights d_k and the auxiliary variables from the model $\widehat{\mathcal{M}}^*_{y,\hat{w}}$ identified in Step 7. In Step 9, the vector of the PMLE of the regression coefficients of the parameters of the final model $\widehat{\mathcal{M}}^*_y$ are adjusted by a matrix $\hat{\mathbf{\Gamma}}_{\mathbf{X}}$ with the PA adjustments (see Section 1.5 for the definition of the PA adjustment). In Step 10, the PA adjusted model $\widehat{\mathcal{M}}_{y,pa}$ is used to produce the PA adjusted fitted means, $\hat{\mu}_{pa,k}$, for the sample. In the last step, the fitted means are substituted into the generic form of the PA estimator, and the estimates of variance are computed using the appropriate formula.

STEP 8. Fit the PML model $\widehat{\mathcal{M}}_y^*$ for y using the auxiliary variables from the model $\widehat{\mathcal{M}}_{y,\hat{w}}^*$ identified in Step 7 using the sampling weight $d_k = \frac{1}{\pi_k}$.

The expressions of the PLL and $\hat{\boldsymbol{\beta}}_{pmle} \in \widehat{\mathcal{M}}_y^*$ are given in (1.3) and (1.4) after replacing $\boldsymbol{\mathcal{M}}_y$ by $\widehat{\mathcal{M}}_y^*$. In this example, the PMLEs of the regression coefficients $\hat{\boldsymbol{\beta}}_{pmle}$ of the models $\widehat{\mathcal{M}}_y^*(y_1)$ and $\widehat{\mathcal{M}}_y^*(y_2)$ for the observed sample are shown in the second column of Table 1.3.

The key outcomes of Step 8 are the auxiliary variables associated with the regression coefficients $\hat{\boldsymbol{\beta}}_{pmle}$ of the working model of the outcome with the best fit.

STEP 9. Create the PA model $\widehat{\mathcal{M}}_{pa,y}$ by adjusting the PMLE of the regression coefficients $\hat{\boldsymbol{\beta}}_{pmle}$ of the model $\widehat{\mathcal{M}}_y^*$ by the PA adjustment $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$.

In this example, because the distribution only includes linear regression coefficients for the location parameter, then the PA adjustment $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}} \in \mathbb{R}^{P \times P}$ is a square matrix where the entries of the main diagonal contain the ratios of the auxiliary variable population total X_k and the HT estimate of the auxiliary population total $\hat{X}_{HT,p}$ for the auxiliary variables $(x_1,\ldots,x_P)$ in the model $\hat{\mathcal{M}}_y^*$ as

$$\hat{\boldsymbol{\Gamma}}_{\mathbf{X}} = \text{diag}\left(\frac{X_1}{\hat{X}_{HT,1}}, ..., \frac{X_P}{\hat{X}_{HT,P}} \right),$$

where $\hat{X}_{HT,p} = \sum_{k \in A} d_k x_{p,k}$ for $p \in \{1, ..., P\}$. The PA adjusted regression coefficients $\hat{\boldsymbol{\beta}}_{pa} \in \mathbb{R}^{P \times 1}$ are

$$\hat{\boldsymbol{\beta}}_{pa} = \hat{\boldsymbol{\Gamma}}_{\mathbf{X}} \hat{\boldsymbol{\beta}}_{pmle} = \begin{pmatrix} \frac{X_1}{\hat{X}_{HT,1}} \hat{\beta}_{pmle,1} \\ ... \\ \frac{X_P}{\hat{X}_{HT,P}} \hat{\beta}_{pmle,P} \end{pmatrix},$$

where $\hat{\boldsymbol{\beta}}_{pmle} = \left(\hat{\beta}_{pmle,1}, ..., \hat{\beta}_{pmle,P} \right)^{\mathrm{T}} \in \mathbb{R}^{P \times 1}$ are the PMLE estimates of the regression coefficients $\boldsymbol{\beta}$ of the model $\widehat{\mathcal{M}}_y^*$. Note that the PA adjustment $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$ is not a calibration adjustment since it does not benchmark the regression coefficients $\hat{\boldsymbol{\beta}}_{pmle}$ to a population total. This step incorporates the information of the population totals into the PLL and the PMLE estimates of the regression coefficients.

The key outcomes of Step 9 are the values of the PA adjusted regression coefficients $\hat{\boldsymbol{\beta}}_{pa}$ of the model $\hat{\mathcal{M}}_y^*$.

The values of $\hat{\boldsymbol{\beta}}_{pmle}$, $\hat{\boldsymbol{\Gamma}}_{pa}$, and the PA adjusted regression coefficient $\hat{\boldsymbol{\beta}}_{pa}$ for the PA models $\widehat{\mathcal{M}}_y^*$ of y_1 and y_2 in this example are shown in the last two columns of Table 1.3.

STEP 10. Estimate the PA adjusted fitted means $\hat{\mu}_{pa,k}$ for the sample cases using the PA model $\hat{\mathcal{M}}_{pa,y}$ from Step 9, and substitute the values $\hat{\mu}_{pa,k}$ into the generic form of the PA estimator for the total Y, $\hat{Y}_{PA} = \sum_{k \in A} d_k \hat{\mu}_{pa,k}$, or the generic form for the mean or proportion $\bar{Y}$, $\hat{\bar{Y}}_{PA} = \dfrac{\hat{Y}_{PA}}{N}$. Then compute the variance estimate of the PA estimator using the appropriate expression (see Section 1.7).

The key outcomes of Step 10 are $\hat{Y}_{PA}$ and $\hat{\mathbb{V}}\left(\hat{Y}_{PA}\right)$ or $\hat{\bar{Y}}_{PA}$ and $\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{PA}\right)$.

Table 1.3 Estimates of the regression coefficient of the model $\widehat{\mathcal{M}}_y^*$ and the PA adjustment of two PA algorithmic estimates in Example 1.1

Model		$\hat{\boldsymbol{\beta}}_{pmle}$	PA adjustment $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$	$\hat{\boldsymbol{\beta}}_{pa}$
y_1: Total hospital expenditures				
Regression coefficient	Auxiliary variable			
$\beta_0 \times 10^{-3}$	1	1,116.04	1.03	1,154.25
$\beta_{11} \times 10^{-3}$	x_{11}	-5,753.23	0.96	-5,515.55
$\beta_2 \times 10^{-3}$	x_2	51.44	1.07	55.05
$\beta_3 \times 10^{-3}$	x_3	166.94	0.79	131.80
$\beta_{52} \times 10^{-3}$	x_{52}	114.94	1.22	140.18
y_2: Indicator of whether hospital received state agency funding				
Regression coefficient	Auxiliary variable			
$\beta_0 \times 10^{3}$	1	34.93	1.03	36.13
$\beta_{12} \times 10^{3}$	x_{12}	204.82	0.93	189.58
$\beta_{15} \times 10^{3}$	x_{15}	965.07	1.08	1,041.57
$\beta_{52} \times 10^{3}$	x_{52}	1.85	1.22	2.25

Since the assumed distribution of y_1 is normal with and identity link function, the PA adjusted fitted mean of y_1, $\hat{\mu}_{pa,y_1,k}$, for the observed sample is

$$\hat{\mu}_{pa,y_1,k} = \hat{\beta}_{pa,0} + \hat{\beta}_{pa,11}\, x_{k11} + \hat{\beta}_{pa,2}\, x_{k2} + \hat{\beta}_{pa,3}\, x_{k3} + \hat{\beta}_{pa,52}\, x_{k52}.$$

Similarly, the PA adjusted fitted mean for y_2, $\hat{\mu}_{pa,y_2,k}$ is

$$\hat{\mu}_{pa,y_1,k} = \hat{\beta}_{pa,0} + \hat{\beta}_{pa,12}\, x_{k12} + \hat{\beta}_{pa,15}\, x_{k15} + \hat{\beta}_{pa,52}\, x_{k52}.$$

The algorithmic PA estimates $\hat{Y}_{PA,1}$ and $\hat{\bar{Y}}_{PA,2}$ for the selected sample listed in Table 1.4 are computed by substituting the PA means $\hat{\mu}_{pa,y_1,k}$ and $\hat{\mu}_{pa,y_2,k}$ in the appropriate generic formula for population total or proportion. The table includes the VDK GREG estimates $\hat{Y}_{VDK,1}$ and $\hat{\bar{Y}}_{VDK,2}$, the estimates of the canonical forms of the HT estimators $\hat{Y}_{HT,1}$ and $\hat{\bar{Y}}_{HT,2}$, and the Hájek (HJ) estimates $\hat{Y}_{HJ,1}$ and $\hat{\bar{Y}}_{HJ,2}$ for reference (see Definition 1.2). The results in Table 1.4 show that for this realization of the sample, for the total Y_1, the relative bias (difference between the estimate and the population value as a percent of the population value) of $\hat{Y}_{PA,1}$ is 17 percent larger than the relative bias of $\hat{Y}_{VDK,1}$. The standard error of $\hat{Y}_{PA,1}$ is 14 percent larger than $\hat{Y}_{VDK,1}$. For the proportion $\bar{Y}_2$, the relative bias of the PA estimate $\hat{Y}_{PA,2}$ is slightly larger than the VDK GREG estimate $\hat{Y}_{VDK,2}$; however, the standard error is 63 percent smaller than the standard error of $\hat{Y}_{VDK,2}$. Although these results are interesting, comparing estimates, bias, and standard errors for one realization is not appropriate for evaluating the performance of the estimators. An

alternative is to compute the same summary statistics under repeated sampling. The empirical statistics for samples of size 80 drawn 100,000 times according to the sample design are summarized in Table 1.5. The table shows the relative bias (RB), relative root mean squared error (RRMSE), the empirical coverage of the 95 percent confidence interval assuming normality, the Kish's design effect (*deff)* (assuming that system of weights are created using the identified working models) and the relative efficiency (RE) with respect to the HT estimator (see the definitions of these empirical summary measures in Section A.4 in Appendix A on page 302).

Table 1.4 Estimates of total Y_1 and proportion $\overline{Y}_2$ based on a single observed sample in Example 1.1

Population characteristic /Estimator	Estimate	Standard error	Kish's design effect (*deff*)	Relative bias (%)
Total Y_1: 8,774,651,373				
$\hat{Y}_{HT,1}$	9,322,853,858	915,126,365	1.31	6.25
$\hat{Y}_{HJ,1}$	9,642,021,099	1,241,508,671	1.31	9.88
$\hat{Y}_{VDK,1}$	9,563,682,688	748,596,001	1.30	8.99
$\hat{Y}_{PA,1}$	9,697,094,833	852,327,681	1.41	10.51
Proportion $\overline{Y}_2$:0.337				
$\hat{\overline{Y}}_{HT,2}$	0.313	0.058	1.31	-7.08
$\hat{\overline{Y}}_{HJ,2}$	0.323	0.059	1.31	-3.90
$\hat{\overline{Y}}_{VDK,2}$	0.340	0.051	1.41	1.07
$\hat{\overline{Y}}_{PA,2}$	0.340	0.032	1. 27	1.09

Table 1.5 shows that all estimators have very small empirical biases as expected even though the working model is misspecified for the binary outcome y_2. The

algorithmic PA estimators $\hat{Y}_{PA,1}$ and $\hat{\bar{Y}}_{PA,2}$ are slightly more efficient than the VDK estimators $\hat{Y}_{VDK,1}$ and $\hat{\bar{Y}}_{VDK,2}$ despite the uncertainty of the model selection in the PA approach. The differences in efficiency between the estimators of Y_1 and $\bar{Y}_2$ are 0.5 and 3.0 percentage points; that is, the PA estimators $\hat{Y}_{PA,1}$ and $\hat{\bar{Y}}_{PA,2}$ are 7.3 percent and 4.0 percent more efficient than the estimators $\hat{Y}_{VDK,1}$ and $\hat{\bar{Y}}_{VDK,2}$, respectively. Furthermore, the expected Kish's design effects of the weights based on the PA estimators are smaller than the design effect of the weights based VDK estimators.

Table 1.5 Empirical summary results* for 100,000 draws for Example 1.1

Population characteristic /Estimator	Relative Bias (RB) (%)	Relative Root Mean Squared Error (RRMSE)	Empirical Coverage of 95% Confidence Interval	Kish's Design effect (*deff*)	Relative efficiency (RE) (%)
Total Y_1					
$\hat{Y}_{HT,1}$	-0.03	9.28	0.946	1.463	0.00
$\hat{Y}_{HJ,1}$	0.60	12.49	0.956	1.463	-44.81
$\hat{Y}_{VDK,1}$	0.58	8.97	0.919	1.502	7.04
$\hat{Y}_{PA,1}$	0.59	8.95	0.911	1.494	7.56
Proportion $\bar{Y}_2$					
$\hat{\bar{Y}}_{HT,2}$	-0.08	19.21	0.935	1.463	0.00
$\hat{\bar{Y}}_{HJ,2}$	0.01	18.04	0.943	1.463	13.40
$\hat{\bar{Y}}_{VDK,2}$	-0.82	14.53	0.923	1.535	74.75
$\hat{\bar{Y}}_{PA,2}$	0.13	14.41	0.924	1.486	77.76

*See Section A.4 in Appendix A for the definitions of the summary measures.

The observed reduction of variance of the GREG and algorithmic PA estimators of Y_1 in Table 1.4 for a single sample is not typical under repeated sampling. In expectation, these estimators are around 7 percent more efficient than the HT estimator. The HT estimator for the total Y_1 is very efficient is due to the high correlation between π_k and y_1. The HJ estimator for the total Y_1, which is also a GREG/PA estimator, is much more inefficient than the HT estimator. Using the population size reduces the efficiency of the HJ estimator of Y_1 considerably.

In contrast, the GREG and PA estimators for the proportion $\bar{Y}_2$ achieve substantial gains of efficiency over the HT estimator, with gains close to 80 percent. The HJ estimator for $\bar{Y}_2$ is around 13 percent more efficient than the HT estimator.

This example shows how the algorithmic PA estimators are developed, and further shows that the algorithmic PA estimators can be more efficient than the VDK GREG estimators based on an in-depth analysis of the full population.

1.4 Principles of the PA Framework

There are four principles of the PA framework that define the roles of working models, auxiliary variable selection, and sample selection.

1. All PA estimators are weighted sums of fitted means of well-defined working models. The fitted mean is a function of linear regressions of auxiliary variables[3] of the parameters of a working model. Different functional models and sets of auxiliary variables yield different PA estimators. Section 3.1 shows that most of the well-known design-based estimators are a subclass of PA estimators.

2. The working models are well defined, but either the functional form or the auxiliary variables of the models (or both) are not known. Most estimators in the survey sampling literature assume the opposite, that is, the functional form is known and the working model is correctly specified (See Deville & Särndal, 1992; Rao, 1994, Lehtonen & Veijanen, 1998; Montanari, 1998; Chen & Sitter, 1999; Wu & Sitter, 2001; Montanari & Ranalli, 2005; Kim, 2009, 2010; Kott, 2016; and Breidt & Opsomer, 2017). Assuming that the working model is correct does not guarantee that the estimator is efficient when the working model is misspecified.

3. The identification of the working model is based on the observed sample. The PA framework produces an estimator that is likely to be efficient based on the sample, but because both the generation of the finite population and the sample selection from it are stochastic processes, there is no guarantee that either the identified working model is the best or that the form of the estimator is unique.

---

[3] The technical definition of this principle is that all estimators are functions of the inverse link function of the linear predictors of the parameters for the location, scale, and shape of the working model. See Definition 1.1 in Section 1.5 for more details.

4. Inferences using the PA estimator are based on the random vector of the sample membership indicator of the element of the population to be selected in the sample. All PA estimators, their variances, and estimates of variances are functions of this random vector; i.e., they are design based. The sequence of a PA estimators for a sequence of increasing population and sample size are model-assisted, so they are asymptotically unbiased and design-consistent under suitable regularity conditions described in Section 5.9.

1.5 Concepts, Definitions, and Notation

The PA framework for estimation with full response assumes two stochastic processes; one is an unobservable process that generates the finite population from a superpopulation model, and the other is based on random sampling from the finite population. Inferences, however, are based only on the random sampling process. In this section, we define the models for these stochastic processes and introduce the notation to facilitate the description of these models in the PA framework. Since a large number of models are defined and evaluated in this approach, we propose a precise notation to describe the working models in the PA framework. We also introduce concepts related to the framework such as the canonical form of an estimator, model misspecification, and valid PA models that are used to describe the PA estimators.

1.5.1 Superpopulation Models

DEFINITION 1.1 Working or assisting model $\mathcal{M}_y$ for the outcome y. Let $\mathcal{M}_y$ be the working model for an outcome y that describes a stochastic process that generates a finite population $\mathcal{F}$ of size N (i.e., $|\mathcal{F}| = N$) as N independent identically distributed (*iid*) realizations from a assumed distribution function f_y defined as

$$y_k \overset{iid}{\sim} \mathrm{f}_Y\left(\boldsymbol{\theta} \mid \mathbf{x}_k\right), \tag{1.5}$$

for $k \in U$, where $\boldsymbol{\theta} \mid \mathbf{x}_k = \left(\theta_\beta, \theta_\sigma, \theta_\gamma\right)^{\mathrm{T}}$ is the vector of the parameters for location, scale, and shape, θ_β, θ_σ, and θ_γ, respectively. We assume that the model parameters are functions of linear predictors of auxiliary variables, then the vector $\boldsymbol{\theta} \mid \mathbf{x}_k$ can be expressed as

$$\boldsymbol{\theta} \mid \mathbf{x}_k = \mathbf{g}^{-1}\left(\boldsymbol{\eta}_k\right) = \begin{pmatrix} \mathrm{g}_\beta^{-1}\left(\eta_{\beta,k}\right) \\ \mathrm{g}_\sigma^{-1}\left(\eta_{\sigma,k}\right) \\ \mathrm{g}_\gamma^{-1}\left(\eta_{\gamma,k}\right) \end{pmatrix}, \tag{1.6}$$

where $\mathbf{g}^{-1}: \mathbb{R}^3 \to \mathbb{R}^3$ is a vector-to-vector function with the inverse of the link functions where g_β, g_σ, and g_γ are the link functions of the parameters for location, scale, and shape, respectively, $\boldsymbol{\eta}_k$ is the vector of the linear predictions $\boldsymbol{\eta}_k = \left(\eta_{\beta,k}, \eta_{\sigma,k}, \eta_{\gamma,k}\right)^{\mathrm{T}}$ with elements defined as

$$\eta_{\beta,k} = \eta_{\beta}\left(\mathbf{x}_{\beta,k}, \boldsymbol{\beta}\right) = \mathbf{x}_{\beta.k}\boldsymbol{\beta}, \tag{1.7}$$
$$\eta_{\sigma,k} = \eta_{\sigma}\left(\mathbf{x}_{\sigma,k}, \boldsymbol{\sigma}\right) = \mathbf{x}_{\sigma,k}\boldsymbol{\sigma}, \text{ and}$$
$$\eta_{\gamma,k} = \eta_{\gamma}\left(\mathbf{x}_{\gamma,k}, \boldsymbol{\gamma}\right) = \mathbf{x}_{k}\boldsymbol{\gamma},$$

where $\eta_{\theta} : \mathbb{R}^{P} \to \mathbb{R}$ is the function $\eta_{\theta}\left(\mathbf{u}, \mathbf{v}\right) = \sum_{p \in P} u_p v_p$ for $\theta \in \{\beta, \sigma, \gamma\}$ where $\mathbf{x}_{\theta,k} \subset \mathbf{x}_k \in \mathbb{R}^{1 \times P}$ are the subset vectors of the auxiliary variable vector $\mathbf{x}_k$ and the parameters $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\gamma}$ are the coefficients of the linear regressions $\eta_{\beta,k}$, $\eta_{\sigma,k}$, and $\eta_{\gamma,k}$, respectively.

REMARK 1.1. In all models, we are interested in the expected value of y_k defined as

$$\mu_k = \mathbb{E}\left(y_k\right) = \int_{R_Y} y_k \, \mathrm{f}_Y\left(y_k\right) dy_k, \tag{1.8}$$

where $R_Y = \left\{ y_k \in \mathbb{R} \mid \mathrm{f}_Y\left(y_k\right) > 0 \right\}$. If the population is available, then the estimate of μ_k is computed by plugging the MLEs of $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\gamma}$ into the expression of $\mu_k = \mathbb{E}\left(y_k\right)$. If only the sample is available, then the estimate of μ_k is computed by plugging the PMLEs of $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\gamma}$ into the expression of μ_k.

The definitions presented above are for the general case. In practice, not all distribution functions have all these parameters defined, as illustrated in the following examples.

EXAMPLE 1.2. Let y be an outcome variable with a distribution $y_k \overset{iid}{\sim} \mathcal{N}\left(\mathbf{x}_k\boldsymbol{\beta}, \sigma_0^2\right)$. This distribution can be described by the vector $\boldsymbol{\theta} \mid \mathbf{x}_k = \left(\theta_\beta, \theta_\sigma\right)^{\mathrm{T}}$ with only two parameters: location and scale. The location parameter is $\theta_\beta \mid \mathbf{x}_k = \eta_{\beta,k}$, the linear predictor is $\eta_{\beta,k} = \mathbf{x}_k\boldsymbol{\beta}$, the vector of auxiliary variables is $\mathbf{x}_k \in \mathbb{R}^{1\times P}$, and the link function is the identity function. The scale parameter is $\theta_\sigma \mid \mathbf{x}_k = \exp\left(\eta_{\sigma,k}\right)$, the linear predictor is $\eta_{\sigma,k} = \sigma_0$, the auxiliary variable is the one vector $\mathbf{1} \in \mathbb{R}^N$, and the link function is $\mathrm{g}_\sigma\left(t\right) = \log\left(t\right)$. Since for this model, $\mu_k = \mathbb{E}\left(y_k\right) = \mathbf{x}_k\boldsymbol{\beta}$, then $\hat{\mu}_{mle,k} = \mathbf{x}_k\hat{\boldsymbol{\beta}}_{mle}$.

EXAMPLE 1.3. Define the outcome y as a log-normal random variable $y_k \overset{iid}{\sim} \log\mathcal{N}\left(\mathbf{x}_k\boldsymbol{\beta}, \mathbf{x}_k^2\sigma^2\right)$. For this distribution, the vector $\boldsymbol{\theta} \mid \mathbf{x}_k = \left(\theta_\beta, \theta_\sigma\right)^{\mathrm{T}}$ contains only the location and scale parameters. The location parameter is $\theta_\beta \mid \mathbf{x}_k = \eta_{\beta,k}$, the linear predictor is $\eta_{\beta,k} = \mathbf{x}_k\boldsymbol{\beta}$, the link function is the identity function, and the auxiliary variables are $\mathbf{x}_k$. The scale parameter is $\theta_\sigma \mid \mathbf{x}_k = \exp\left(\eta_{\sigma,k}\right)$, the linear predictor is $\eta_{\sigma,k2} = \mathbf{x}_k\boldsymbol{\sigma}$, the link function is $\mathrm{g}_\sigma\left(t\right) = \log\left(t\right)$, and the auxiliary variables are $\mathbf{x}_k$. Since for the lognormal distribution $\mu_k = \exp\left(\theta_\beta + \frac{\theta_\sigma^2}{2}\right)$, then $\hat{\mu}_{mle,k} = \exp\left(\mathbf{x}_k\hat{\boldsymbol{\beta}}_{mle} + \frac{\left(\mathbf{x}_k\hat{\boldsymbol{\sigma}}_{mle}\right)^2}{2}\right)$.

EXAMPLE 1.4. In the example in Section 2.4 on page 152, the outcome y is assumed to be normally distributed $y_k \mid \mathbf{x}_k \overset{iid}{\sim} \mathcal{N}\left(\theta_{\beta,k}, \theta_{\sigma,k}^2\right)$ where

$$\begin{aligned} \theta_{\beta,k} &= \mu_k = \beta_0 + \beta_1 \pi_k + \beta_2 d_k \text{ and} \\ \theta_{\sigma,k}^2 &= \left(\exp\left(\sigma_0 + \sigma_1 \pi_k + \sigma_2 d_k\right)\right)^2 \left|\mu_k\right|^{\gamma_0}. \end{aligned} \tag{1.9}$$

The model of y_k is appropriate for normally distributed regression models where the variance of the response variable is proportional to a power of the mean. The auxiliary variables are $\mathbf{x}_k = \left(1, \pi_k, d_k\right)$. The elements of the vector $\boldsymbol{\theta} \mid \mathbf{x}_k = \left(\theta_\beta, \theta_\sigma, \theta_\gamma\right)^{\mathrm{T}}$ are the location parameter $\theta_\beta \mid \mathbf{x}_k = \eta_{\beta,k}$ with a linear predictor $\eta_{\beta,k} = \beta_0 + \beta_1 \pi_k + \beta_2 d_k$, the link function is the identity; the scale parameter is $\theta_\sigma \mid \mathbf{x}_k = \exp\left(\eta_{\sigma,k}\right)$ with a linear predictor $\eta_{\sigma,k} = \sigma_0 + \sigma_1 \pi_k + \sigma_2 d_k$ with the link function $\mathrm{g}_\sigma\left(t\right) = \log\left(t\right)$, the shape parameter $\theta_\gamma \mid \mathbf{x}_k = \eta_{\gamma,k}$ with a linear predictor $\eta_{\gamma,k} = \gamma_0$, and the link function is the identity. Since for this model $\mu_k = \beta_0 + \beta_1 \pi_k + \beta_2 d_k$, then $\hat{\mu}_{mle,k} = \hat{\beta}_{mle,0} + \hat{\beta}_{mle,1} \pi_k + \hat{\beta}_{mle,2} d_k$.

REMARK 1.2. The parametric models described in the expressions (1.5), (1.6), and (1.7) are a subset of the models known as generalized additive models for location, scale, and shape (GAMLSS) proposed by Stasinopoulos et al. (2017). The GAMLSS is an extension of the GLM proposed by McCullagh & Nelder (1989)[4]. In

[4]A similar extension of the GLM is the Vector generalized linear model (VGLM) proposed by Yee (2015).

the GLMs, only the location parameter θ is a function of the linear regression of the auxiliary variables, but in GAMLSS, the location, scale, and shape parameters are also modeled using linear combinations of auxiliary variables and link functions. The GAMLSS allows distributions where $\mathbb{E}(y) = \mu \neq \mathrm{g}^{-1}(\mathbf{x}\boldsymbol{\beta})$ such as the lognormal and zero-inflated Poisson. Although the GAMLS includes a large number of models, most models we study include just a location parameter and, in a few instances, a scale parameter.

DEFINITION 1.2 Working model $\mathcal{M}_\pi$ for the sample membership indicator $\mathbf{S} = \mathbf{s}$. In the PA framework for estimation with full response, we assume working models for $\mathbf{S}$ that do not need to be correctly specified because these models are only used to identify explanatory auxiliary variables of $\mathbf{S}$. These models are approximations of the sample design in Definition 1.5. The definition of $\mathcal{M}_\pi$ is similar to the definition of the outcome working model $\mathcal{M}_y$ described above. The model $\mathcal{M}_\pi$ may have simpler distributions without separate scale and location parameters. The probability mass function of $\mathcal{M}_\pi$ for a random vector of the sample membership indicator $S_k = s_k$ (e.g., $s_k = 1$ if the element k was selected in the sample or $s_k = 0$ otherwise) is generally modeled using the Bernoulli distribution $\mathcal{B}e(\pi_k)$ where $\Pr(S_k = s_k = 1 \mid \mathbf{x}_k) = \pi_k$ and $\Pr(S_k = s_k = 0 \mid \mathbf{x}_k) = 1 - \pi_k$ with link functions such as the logit model, $\mathrm{logit}(\pi_k) = \mathbf{x}_k\boldsymbol{\beta}$ or the linear probability model, $\pi_k = \mathbf{x}_k\boldsymbol{\beta}$, (Cox, 1970). Finding the MLE of the parameters of the working

model $\mathcal{M}_\pi$ requires access to the entire population. If this is the case, the parameters $\boldsymbol{\beta}$ are estimated using logistic regression with $s_k \in \{0,1\}$ as the dependent variable. Then $\hat{\pi}_k$ is computed by plugging the estimates $\hat{\boldsymbol{\beta}}$ into the formula for $\mathbb{E}(S_k)$. If only the sample is available, then the model $\mathcal{M}_\pi$ is fitted using PL logistic regression using the sampling weight d_k.

REMARK 1.3. An alternative for modeling S_k, the random variable with the membership sample indicator, is directly modeling the inclusion probability assuming that π_k for $k \in U$ are the realizations of a random variable from the superpopulation model $\pi_k \sim \mathrm{f}(\boldsymbol{\theta} \mid \mathbf{x}_k)$. Some distributions for the working models for $\pi_k \in (0,1)$ are:

1. The beta distribution $\mathcal{B}eta(\alpha, \beta)$ with location parameter $\theta_\beta = \dfrac{\alpha}{\alpha + \beta}$ and scale parameter $\theta_\sigma = \dfrac{1}{\sqrt{\alpha + \beta + 1}}$, where $\pi_k = \mathbb{E}(\theta_\beta \mid \mathbf{x}_k) = \mathrm{logit}^{-1}(\mathbf{x}_k \boldsymbol{\beta})$ (Ferrar & Cribari-Neto, 2004). The regression coefficients $\hat{\boldsymbol{\beta}}_{mle}$ in $\hat{\pi}_k = \mathrm{logit}^{-1}(\mathbf{x}_k \hat{\boldsymbol{\beta}}_{mle})$ are computed using GLM beta regression using the entire population (Stasinopoulos, Rigby, Heller, Voudouris, & De Bastiani, 2017).

2. The "fractional logit" model for fractional response variables $\pi_k \in (0,1)$ (Papke & Wooldridge, 1996). The parameters $\hat{\boldsymbol{\beta}}$ in $\hat{\pi}_k = \mathrm{logit}^{-1}(\mathbf{x}_k \hat{\boldsymbol{\beta}})$ are computed using quasi-maximum likelihood (QL) with π_k as the dependent variable

(Wedderburn, 1974). The QL estimators are used when the form of distribution is unknown but can be approximated by the mean and variance. Although the QL is related to the likelihood, it is not the same since the exact distribution is not known. A quasi-maximum likelihood estimator (QMLE) of the parameter θ of a model is computed by maximizing the QL. Finite sample properties of QMLE and QL have not been fully studied in survey sampling although they are currently used in practice (see Lumley, 2010).

3. A misspecified Bernoulli distribution with the dependent variable $\pi_k \in (0,1)$ computed as $\hat{\pi}_k = \text{logit}^{-1}\left(\mathbf{x}_k \hat{\boldsymbol{\beta}}_{qmle}\right)$. Strictly speaking, the Bernoulli distribution for π_k is misspecified because the support of the distribution is $\{0,1\}$ while π_k takes fractional values between zero and one. However, Gourieroux, Monfort, & Trognon (1984) show that the MLEs of the parameters of misspecified models with a distribution from the linear exponential family are consistent estimates of the MLE parameters of any other linear exponential family distribution including the parameters of the correct model. These results justify the use of both the logistic regression and the linear probability model for the fractional values of π.

4. The linear probability model $\mathcal{N}\left(\theta_\beta, \theta_\sigma^2\right)$ for π_k where $\mathbb{E}(\pi_k) = \theta_\beta = \mathbf{x}_k \boldsymbol{\beta}$ and $\theta_\sigma = \sigma$ (Greene, 2008). The estimated MLE parameters $\hat{\boldsymbol{\beta}}_{mle}$ in $\hat{\pi}_k = \mathbf{x}_k \hat{\boldsymbol{\beta}}_{mle}$ are computed using linear regression. This model is misspecified since the values of $\hat{\pi}_k$ may be outside the support of π_k.

5. Any other distribution that fits the shape of π_k, for example, the logistic distribution

$$\mathrm{f}\left(\pi \mid \theta_\beta, \theta_\sigma\right) = \frac{1}{\theta_\sigma} \frac{\exp\left(-\dfrac{\pi - \theta_\beta}{\theta_\sigma}\right)}{\left(1 + \exp\left(-\dfrac{\pi - \theta_\beta}{\theta_\sigma}\right)\right)^2},$$

where $\mathbb{E}\left(\pi_k\right) = \theta_\beta = x_k\boldsymbol{\beta}$ and $\theta_\sigma = \sigma$ (Johnson, Kotz, & Balakrishnan, 1994). The MLE parameters $\hat{\boldsymbol{\beta}}_{mle}$ in $\hat{\pi}_k = \mathbf{x}_k\hat{\boldsymbol{\beta}}_{mle}$ are computed using GAMLSS regression (Stasinopoulos, Rigby, Voudouris, Akantziliotou, Enea, Kiose, 2017).

REMARK 1.4. If only the sample is available and π_k for $k \in A$ are known, then the model $\mathcal{M}_\pi$ is fitted using PLL and the sampling weights d_k. See Section 1.6 for the empirical properties of algorithmic PA estimators that directly model the probabilities of selection π for the population and sample design in the example in Section 1.3 on page 6.

REMARK 1.5. Beaumont (2008) proposes a method to improve the efficiency of the estimators by smoothing design or calibration weights using an appropriate model. His method produces a single set of smoothed weights for multipurpose surveys with estimators $\hat{Y}_B = \sum_{k \in A} \hat{\tilde{w}}_k y_k$ where $\hat{\tilde{w}}_k$ is the estimated smoothed weight. This approach differs from the PA algorithm that models the sample membership indicators and uses the fitted means of the working model of the

probabilities of inclusion to produce adjusted weights, but these adjusted weights are not used in to create the PA estimator $\hat{Y}_{PA} = \sum_{k \in A} d_k \hat{\mu}_{pa,k}$ in Step 10. The PA estimator can be seen as an estimator with improved efficiency that results from smoothing the outcome variable y.

There are other differences between the two approaches. For example, Beaumont (2008) states that any classical model selection and validation techniques can be used to determine an appropriate model and does not use the design weights in the modeling. Furthermore, the smoothed-weight estimators can be biased as shown in his simulation study, while the PA estimators are design consistent with small bias, even in relatively small samples.

DEFINITION 1.3 The collection of working models $\mathcal{M}_y$ for the outcome variable y. Let $\mathcal{M}_y$ be the collection of, at most, three sets of working models for the scale, location, and shape of the distribution of y denoted as

$$\mathcal{M}_y = \mathcal{M}_y\left(\theta_\beta\right) \cup \mathcal{M}_y\left(\theta_\sigma\right) \cup \mathcal{M}_y\left(\theta_\gamma\right), \tag{1.10}$$

where each set of models $\mathcal{M}_y\left(\theta\right)$ for $\theta \in \left\{\theta_\beta, \theta_\sigma, \theta_\gamma\right\}$ is defined as

$$\mathcal{M}_y\left(\theta\right) = \text{span}\left(\mathbf{x}\right) = \left\{ \sum_{p=1}^{P} \theta_p x_{\theta p} \mid P \in \mathbb{N}, x_{\theta p} \in \mathbf{x}, \theta_p \in \Theta_{PA} \right\}, \tag{1.11}$$

where $\mathbf{x}_\theta \in \mathbb{R}^{p_\theta}$ such as $\mathbf{x}_\theta \subseteq \mathbf{x}$ is the set of auxiliary variables associated with the parameter θ. Each $\mathcal{M}_y\left(\theta\right)$ is a spanned subspace with all linear combinations of

the auxiliary variables $\mathbf{x}_\theta$ and the parameters θ_p for $p \in \{1,..,P_\theta\}$ that produce a valid PA model in the set Θ_{PA}. In other words, by definition, the vector space with all models generated by the vector $\mathbf{x}$ excludes the invalid PA models (see Definition 1.19). Note that despite the finite number of linear combinations, there is an infinite number of models in $\mathcal{M}_y$ because the parameters can take any valid value in their support depending on the distribution f_Y of y and link functions.

The collection of working models $\mathcal{M}_\pi$ for sample membership indicator $\mathbf{S}$ or for the inclusion probability π is defined the same way as $\mathcal{M}_y$.

1.5.2 Notation for the Collection of Models $\mathcal{M}_y$

Since the number of models described by $\mathcal{M}_y$ or $\mathcal{M}_\pi$ is large, we need a precise notation for describing the model. Since the spanned set of models in (1.11) includes the models formed by linear combinations of the auxiliary variables for the parameters $\theta \in \{\theta_\beta, \theta_\sigma, \theta_\gamma\}$, then each model in the collection $\mathcal{M}_y(\theta)$ can be uniquely identified by the auxiliary variables or parameters of linear predictions $\hat{\eta}_\beta$, $\hat{\eta}_\sigma$, and $\hat{\eta}_\gamma$. Based on this idea, we can use two notations for identifying these models as illustrated in the following examples.

EXAMPLE 1.5. The collection of models $\mathcal{M}_y$ for the outcome y. Let y_k be a random variable assumed to follow a normal distribution $y \overset{iid}{\sim} \mathcal{N}\left(\theta_\beta, \theta_\sigma^{2\theta_\gamma}\right)$ where $\theta_\beta = \beta_0 + \beta_1 x_{k1}$, $\theta_\sigma = \sigma_0 + \sigma_2 x_k$, and $\theta_\gamma = \gamma_0 + \gamma_4 x_{k4}$, with the vector of auxiliary variables $\mathbf{x} = (1, x_2, x_3, x_4)$. The first notation or full notation of all possible models in the collection of models $\mathcal{M}_y$ uses the matrix $\left(\boldsymbol{\theta}_\beta^{\mathrm{T}}, \boldsymbol{\theta}_\sigma^{\mathrm{T}}, \boldsymbol{\theta}_\gamma^{\mathrm{T}}\right)$ with model membership indicators for the regression coefficients using the position of the associated variable in the vector of the auxiliary variables $\mathbf{x}$ as shown in the fourth column of Table 1.6. For this example, the full notation for the collection of models is the matrix

$$\mathcal{M}_y = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \tag{1.12}$$

where the entries of the rows of the matrix with values of one indicate the variables that appear in the linear predictors of the location parameter (first row), scale parameter (second row), and shape parameter (third row) of the model $\mathcal{M}_y$. If the auxiliary variable does not appear in the linear predictor, the entry has a value of zero. The main disadvantage of the full notation is that the order of the auxiliary variables in the vector $\mathbf{x}$ needs to be known. Furthermore, as the number of auxiliary variables increases (e.g., including dummy indicators for each level of categorical variables or variables for interaction terms), the matrix $\mathcal{M}_y$ becomes difficult to read.

We propose a simplified notation or short notation that only lists either the nonzero regression coefficients or their associated auxiliary variables in each parameter model, as shown in the last two columns of Table 1.6. Using the short notation, the collection of models $\mathcal{M}_y$ in Example 1.5 is either $\mathcal{M}_y = \{(1, x_1), (1, x_2), (1, x_3)\}$ or $\mathcal{M}_y = \{(\beta_0, \beta_1), (\sigma_0, \sigma_2), (\gamma_0, \gamma_3)\}$. We prefer to list the auxiliary variables of the models because the values of the regression coefficients are not relevant except for the models fitted in the last steps of the algorithm.

Table 1.6 Full and Simplified Notations for the collection of models $\mathcal{M}_y$ for Example 1.2

Collection of models	Model parameter	Linear prediction (η)	Model notation: Full: Membership indicators based on $\mathbf{x} = (1, x_1, x_2, x_3)$	Model notation: Simplified: Regression coefficients	Model notation: Simplified: Auxiliary variables
$\mathcal{M}_y(\theta_\beta)$	Location	$\beta_0 + \beta_1 x_{k1}$	$(1,1,0,0)$	(β_0, β_1)	$(1, x_1)$
$\mathcal{M}_y(\theta_\sigma)$	Scale	$\sigma_0 + \sigma_2 x_{k2}$	$(1,0,1,0)$	(σ_0, σ_2)	$(1, x_2)$
$\mathcal{M}_y(\theta_\gamma)$	Shape	$\gamma_0 + \gamma_3 x_{k3}$	$(1,0,0,1)$	(γ_0, γ_2)	$(1, x_3)$

If we extend the short notation, then the auxiliary variables for categorical variables are written in boldface since they represent a vector of membership indicators (e.g., dummy variables with one and zero values) for each categorical level. For interaction terms, we write the product of the two variables. For example, suppose that there are sampling stratum indicators $\mathbf{h}_k = (h_{k1}, ..., h_{kh'}, ..., h_{kH})$ for $h' \in \{1, ..., H\}$ where $h_{kh'} = 1$ if the element k belongs to stratum h' and zero otherwise, and H is the number of strata. If we want to describe the collection of models where the linear

predictor for the location parameter η_β includes the sampling stratum indicators $\mathbf{h}_k$ and the interaction between the x_1 and $\mathbf{h}_k$, then the collection of models for y is written as

$$\mathcal{M}_y = \left\{(1, x_1, \mathbf{h}, \mathbf{h} * x_1), (1, x_2), (1, x_3)\right\},$$

where $\mathbf{h}_k * x_{k1} = \left(h_{k1}x_{k1}, ..., h_{kh'}x_{k1}, ..., h_{kH}x_{k1}\right)$ for $k \in U$.

The short notation can be further simplified by including only the auxiliary variables of the model parameters used to compute $\hat{\mu}_k$. Returning to the example, since y_k is assumed to be normally distributed, then the short notation of the model only includes the auxiliary variables of location parameter as $\mathcal{M}_y = (1, x_1, \mathbf{h}, \mathbf{h} * x_1)$.

EXAMPLE 1.6. Let S_k be the random variable for the sample membership indicator for a stratified design with two strata with indicators $\mathbf{h}_k = (h_{k1}, h_{k2})$ and one continuous auxiliary variable x_2. We assume that the distribution of S_k is $S_k \overset{iid}{\sim} \mathcal{B}e(\pi_k)$ with a link function $\text{logit}(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right)$. Using the simplified notation, the collection of models for π is $\mathcal{M}_\pi = (1, \mathbf{h}, x_2, \mathbf{h} * x_2)$ or $\mathcal{M}_\pi = (1, \mathbf{h}, x_2, \mathbf{x}_3)$ where $\mathbf{x}_3$ is the vector for the interaction terms between $\mathbf{h}$ and x_2, defined as $\mathbf{x}_{k3} = \mathbf{h}_k * x_{k2} = (\text{h}_{k1}\, x_{k2}, \text{h}_{k2}\, x_{k2})$. In this case, since the distribution of y does not have a shape parameter and the scale parameter is a function of the location parameter, there is no need to include these parameters in $\mathcal{M}_\pi$.

1.5.3 Finite Populations and Sample Designs

The following definitions are related to the finite population assumed to be N *iid* realizations of a superpopulation model.

DEFINITION 1.4. Finite population $\mathcal{F}$. We follow the Fuller (2009) notation. Let $U=\{1,\ldots,N\}$ be the labels identifying each element of a finite population of known size N. Associated with the element $k\in U$ is a row data vector $(y_k,\mathbf{x}_k)\in\mathbb{R}^{1\times(P+1)}$ where $y_k\in\mathbb{R}$ is the study variable and $\mathbf{x}_k\in\mathbb{R}^{1\times P}$ is the vector of the auxiliary variables $\mathbf{x}_k=\left[x_{kp}\right]=(x_{k1},...,x_{kP})$ with $P\in\mathbb{N}$, $p\in\{1,...,P\}$, and $P\ll N$. The finite population is defined as the entire set $\mathcal{F}=\{(y_1,\mathbf{x}_1),(y_2,\mathbf{x}_2),...,(y_N,\mathbf{x}_N)\}$, which is assumed to be generated by a working model $\mathcal{M}_y$. We assume that population totals denoted by $\mathbf{X}\in\mathbb{R}^{1\times P}$ where $\mathbf{X}=\sum_{k\in U}\mathbf{x}_k=(X_1,\ldots,X_P)$ of the auxiliary variables $\mathbf{x}_k$ are known.

DEFINITION 1.5. Model for the sample design $p(A=a)$ (see Fuller, 2009). Let A be a subset of U and let $\mathcal{A}$ be the collection of subsets of U that contains all possible samples. Let $\Pr(A=a)$ denote the probability that a, $a\in\mathcal{A}$, is selected. A sampling design is the function that maps the event that $a\in\mathcal{A}$ is selected to $[0,1]$ such that $p(a)=\Pr(A=a)$ for any $a\in\mathcal{A}$. Let π_k be the first-order inclusion

probabilities for element $k \in U$ where $\pi_k = \Pr(k \in A) = \sum_{a \in A_{(i)}} p(a)$ and $A_{(k)}$ is the set of samples that contain the element k. In this dissertation, we consider only single-stage, without replacement sample designs.

Let $\mathbf{S} = (S_1, \ldots, S_N)^{\mathrm{T}} \in \{0,1\}^N$ be a vector of discreet random variables for the sample membership indicators, $S_k = \{0,1\}$, for all elements of the frame where s_k is the realization of S_k defined as

$$S_k = s_k = \begin{cases} 1 & \text{if unit } k \text{ is selected in the sample} \\ 0 & \text{Otherwise} \end{cases}. \tag{1.13}$$

The sample design determines the probability structure of $\mathbf{S}$ that determines the probability behavior of functions of the sample for $k \in U$. Let π_k be the first order inclusion probability of unit k defined as $\mathbb{E}(S_k \mid \mathcal{F}) = \pi_k \in (0,1)$. We use A as the set of indices subset of U that appear in the sample. The (observed) sample size is defined as $n_o = \sum_{k \in U} s_k = \sum_{k \in A} s_k$.

Using the Tillé (2006) notation, the sample design is defined by a random vector $\mathbf{S} \in \{0,1\}^N$ with discrete random variables, S_k, that follows as a multinomial distribution with an expected value $\mathbb{E}(\mathbf{S} \mid \mathcal{F}) = \boldsymbol{\pi} = [\pi_k] \in (0,1)^N$ where $\boldsymbol{\pi}$ is the vector of the probabilities of inclusion π_k for $k \in U$ and the variance-covariance matrix of $\mathbf{S}$, $\boldsymbol{\Delta}$ defined as

$$\begin{aligned}\boldsymbol{\Lambda} &= \mathbb{C}\left(\mathbf{S} \mid \mathcal{F}\right) \\ &= \mathbb{E}\left(\mathbf{S}\mathbf{S}^{\mathrm{T}} \mid \mathcal{F}\right) - \mathbb{E}\left(\mathbf{S} \mid \mathcal{F}\right)\mathbb{E}\left(\mathbf{S}^{\mathrm{T}} \mid \mathcal{F}\right), \\ &= \boldsymbol{\Pi} - \boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}}\end{aligned} \tag{1.14}$$

where $\boldsymbol{\Pi} = \left[\pi_{kl}\right] \in \mathbb{R}^{N \times N}$ is the matrix with the second-order probability of inclusion π_{kl} of units k and l defined as the probability the 2-tuple $\left(k,l\right)$ is selected in the sample at the same time, $\pi_{kl} = \mathbb{E}\left(S_k, S_l \mid \mathcal{F}\right)$ for $k \neq l \in U$ or $\pi_{kk} = \pi_k$ for $k = l \in U$. In matrix notation, the population $\mathcal{F}$ or frame is the matrix $\left(\mathbf{y}, \mathbf{x}\right)$, and the matrix of auxiliary variables $\mathbf{x}$ is the design matrix. The observed data in the sample correspond to the matrices $\left(\mathbf{y} \odot \mathbf{S}, \mathbf{x}\right)$ or $\left(\mathbf{y} \odot \mathbf{S}, \mathbf{x} \odot \mathbf{S}\right)$, the latter if the values of $\mathbf{x}_k$ are only observed in the sample. The operator $\odot$ is the Hadamard-Schur or element-wise matrix product (Horn & Johnson, 2013). The expected sample size is $n = \mathbf{1}^{\mathrm{T}}\boldsymbol{\pi}$, and the variance of the sample size is $\mathbb{V}\left(n \mid \mathcal{F}\right) = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Delta}\mathbf{1}$.

DEFINITION 1.6. Sample designs where the variance of the sample size $\mathbb{V}\left(n \mid \mathcal{F}\right) = 0$ are called fixed size or fixed sample size designs. Those designs that do not meet this condition are called variable size, random size, or random sample size designs.

1.5.4 The Log-Likelihood and Pseudo-Likelihood

DEFINITION 1.7. The log-likelihood function and the maximum likelihood estimators of the working model fitted to the full population. The expression of the LL of the model $\mathcal{M}_y$ for the variable y fitted to the finite population U (e.g., census fit) is

$$\log \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{x}) = \sum_{k \in U} \log \mathrm{f}_Y(y_k, \mathbf{x}_k \mid \boldsymbol{\theta}). \tag{1.15}$$

The MLE of $\boldsymbol{\theta}$ is computed as

$$\hat{\boldsymbol{\theta}}_{mle} \in \left\{ \underset{\theta \in \Theta}{\arg\max} \log \mathcal{L}(\boldsymbol{\theta}) \right\}. \tag{1.16}$$

See Cheng (2017) for the regularity conditions for the asymptotic properties of the MLEs. Under these regularity conditions, the MLE $\hat{\boldsymbol{\theta}}_{mle}$ exists and is unique. A similar expression is available for the MLE of the sample membership indicator $\mathbf{S}$ (see Section 1.6 for models for π).

DEFINITION 1.8. The collection of ML working models $\widehat{\boldsymbol{\mathcal{M}}}_{mle,y}$ for the outcome variable y. Let $\widehat{\mathcal{M}}_{mle,y} \subseteq \mathcal{M}_y$ be the collection of MLE models of y defined as the subset of the models in $\boldsymbol{\mathcal{M}}_y$, where the estimates of the regression coefficients of the parameters are MLEs. Using the simplified notation, $\widehat{\mathcal{M}}_{mle,y} = \left\{ \hat{\boldsymbol{\beta}}_{mle,\mathbf{x}_\beta}, \hat{\boldsymbol{\sigma}}_{mle,\mathbf{x}_\sigma}, \hat{\boldsymbol{\gamma}}_{mle,\mathbf{x}_\gamma} \right\}$, where $\hat{\boldsymbol{\beta}}_{mle,\mathbf{x}_\beta}$, $\hat{\boldsymbol{\sigma}}_{mle,\mathbf{x}_\sigma}$, and $\hat{\boldsymbol{\gamma}}_{mle,\mathbf{x}_\gamma}$ are the

MLEs of the location parameters $\boldsymbol{\beta}$, scale parameters $\boldsymbol{\sigma}$, and shape parameters $\boldsymbol{\gamma}$ of the models in $\boldsymbol{\mathcal{M}}_y$ and $\mathbf{x}_\theta \subseteq \mathbf{x}$ for $\theta \in \{\beta, \sigma, \gamma\}$ are the subsets of the auxiliary variables $\mathbf{x}$ for the location, scale, and shape parameters. The auxiliary variables $\mathbf{x}_\theta$ are not necessarily the same for the location, scale, and shape parameters in $\boldsymbol{\mathcal{M}}_y$. See Section 1.6 for the method for generating the models in $\boldsymbol{\mathcal{M}}_y$ and computing the models $\widehat{\boldsymbol{\mathcal{M}}}_{mle,y}$. The collection of ML models for $\mathbf{S}$, $\widehat{\boldsymbol{\mathcal{M}}}_{mle,\pi}$, has a similar expression as $\widehat{\boldsymbol{\mathcal{M}}}_{mle,y}$. For notation convenience, we drop the subscripts of the auxiliary variables of the parameters with the understanding that different subsets of auxiliary variables are associated with these parameters.

DEFINITION 1.9. The best-fit ML model $\widehat{\mathcal{M}}_y \in \widehat{\boldsymbol{\mathcal{M}}}_y$ for y. All models in $\widehat{\boldsymbol{\mathcal{M}}}_y$ are created using the MLE $\hat{\boldsymbol{\theta}}_{mle} = \left(\hat{\boldsymbol{\beta}}_{mle}, \hat{\boldsymbol{\sigma}}_{mle}, \hat{\boldsymbol{\gamma}}_{mle}\right)^{\mathrm{T}}$; however, some ML models have a better fit to the observed sample than others. The ML models in $\widehat{\boldsymbol{\mathcal{M}}}_y$ can be ranked based on the values of a loss function $\mathrm{L}(y)$ that measures goodness of fit of the models. Let $\widehat{\mathcal{M}}_y \in \widehat{\boldsymbol{\mathcal{M}}}_y$ be the ML model that achieves the lowest value of the loss function $\mathrm{L}\left(\widehat{\mathcal{M}}_y\right)$ (see Section 1.6 for the definition of the loss function and how the model in $\widehat{\mathcal{M}}_y$ is found among the models in $\widehat{\mathcal{M}}_y$). The MLE of $\mu_k = \mathbb{E}(y_k)$ is obtained by plugging the ML estimates $\hat{\boldsymbol{\beta}}_{mle}$, $\hat{\boldsymbol{\sigma}}_{mle}$, and $\hat{\boldsymbol{\gamma}}_{pmle}$ into the expression of μ_k of the specific distribution of the working model. The

expressions of the collection of ML models $\widehat{\boldsymbol{\mathcal{M}}}_{\pi}$ and the best-fit model $\widehat{\mathcal{M}}_{\pi}$ for $\mathbf{S}$ is similar to the expressions of $\widehat{\boldsymbol{\mathcal{M}}}_{y}$ and $\widehat{\mathcal{M}}_{y}$ for y.

DEFINITION 1.10. The pseudo-log-likelihood function and the pseudo-maximum likelihood estimators fitted to the sample. The PL of the model $\mathcal{M}_{y}$ for y of fitted to sample A is defined as

$$\log \mathcal{L}_A\left(\boldsymbol{\theta}; \mathbf{y}, \mathbf{x}, \mathbf{d} \mid \mathcal{F}\right) = \sum_{k \in A} d_k \log \mathrm{f}_Y\left(y_k, \mathbf{x}_k \mid \boldsymbol{\theta}\right), \tag{1.17}$$

where $\mathbf{d} = \left[d_k\right]$ are the sampling weights for $k \in A$. The PMLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_{pmle} \in \left\{ \arg\max_{\theta \in \Theta} \log \mathcal{L}_A\left(\boldsymbol{\theta}\right) \right\}. \tag{1.18}$$

See Binder (1983) for the regularity conditions for the asymptotic properties of the PMLEs. Under these conditions, the PML estimate $\hat{\boldsymbol{\theta}}_{pmle}$ exists and is unique. A similar expression is available for the PL and PMLE of the sample membership indicator $\mathbf{S}$ and the inclusion probability π.

DEFINITION 1.11. The collection of PML models $\widehat{\boldsymbol{\mathcal{M}}}_{y}$ for y. The collection of PML models, $\widehat{\boldsymbol{\mathcal{M}}}_{y}$ is defined in the same way as the ML models for y, but replacing the MLEs of $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\gamma}$ by the corresponding PMLEs. Using the simplified notation, $\widehat{\boldsymbol{\mathcal{M}}}_{y} = \left(\hat{\boldsymbol{\beta}}_{pmle,\mathbf{x}_\beta}, \hat{\boldsymbol{\sigma}}_{pmle,\mathbf{x}_\sigma}, \hat{\boldsymbol{\gamma}}_{pmle,\mathbf{x}_\gamma}\right)$ where $\hat{\boldsymbol{\beta}}_{pmle,\mathbf{x}_\beta}$, $\hat{\boldsymbol{\sigma}}_{pmle,\mathbf{x}_\sigma}$,

and $\hat{\boldsymbol{\gamma}}_{pmle,\mathbf{x}_\gamma}$ are the PMLEs of the location parameters $\boldsymbol{\beta}$, scale parameters $\boldsymbol{\sigma}$, and shape parameters $\boldsymbol{\gamma}$ of the models in $\boldsymbol{\mathcal{M}}_y$ with the auxiliary variables $\mathbf{x}_\theta \subseteq \mathbf{x}$ for $\theta \in \{\beta, \sigma, \gamma\}$. The auxiliary variables $\mathbf{x}_\theta$ are not necessarily the same for the location, scale, and shape parameters in $\boldsymbol{\mathcal{M}}_y$. The collection of PML models for $\mathbf{S}$ (or π), $\widehat{\boldsymbol{\mathcal{M}}}_\pi$, has an expression similar to the models $\widehat{\boldsymbol{\mathcal{M}}}_y$ for y.

DEFINITION 1.12. The best fit PML model $\widehat{\mathcal{M}}_y \in \widehat{\boldsymbol{\mathcal{M}}}_y$ for y. The best fit PML model $\widehat{\mathcal{M}}_y \in \widehat{\boldsymbol{\mathcal{M}}}_y$ for y is defined in the same way as the ML model in Definition 5.3 but using a loss function $\mathrm{L}\left(\widehat{\boldsymbol{\mathcal{M}}}_y\right)$ based on the sample estimate of the goodness of fit of the PML model. In the current implementation of the PA approach, we use the sample-based AIC as dAIC. See Section A.2 in Appendix A for details on the dAIC.

REMARK 1.6. We assume that the finite population is a realization of a superpopulation model $\mathcal{M}_y$; however, the parameters and their values are unknown (See Principle 2 in Section 1.4). When identifying the superpopulation model, we need to determine its functional form and the parameters (and their associated auxiliary variables). See Definition 1.1 and Principle 1 in Section 1.4. When the entire population is analyzed, multiple sets of MLEs $\left\{\hat{\boldsymbol{\beta}}_{mle}, \hat{\boldsymbol{\sigma}}_{mle,}, \hat{\boldsymbol{\gamma}}_{mle}\right\} \in \widehat{\boldsymbol{\mathcal{M}}}_{mle,y}$ can be fitted to the population data $\mathcal{F}$ since they are formed by the combinations of the

parameters and auxiliary variables. These sets of MLE of the regression coefficients are efficient and consistent estimators of their corresponding regression coefficients, $\{\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{\gamma}\} \in \boldsymbol{\mathcal{M}}_y$ of the superpopulation models (assuming that each model is the true model). To identify a single model among all ML models, we use the goodness of fit; that is, we assume that the true superpopulation model has the lowest discrepancy between the observed population values and the expected values from the fitted model as measured by a loss function. We denote the best-fit ML model as $\widehat{\mathcal{M}}_y$ where the MLEs of the regression coefficients $\left(\hat{\boldsymbol{\beta}}_{mle},\hat{\boldsymbol{\sigma}}_{mle},\hat{\boldsymbol{\gamma}}_{mle}\right) \in \widehat{\mathcal{M}}_y$ are efficient and consistent estimators of the regression coefficients $(\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{\gamma}) \in \mathcal{M}_y$ and $\mathcal{M}_y$ is the assumed true superpopulation model.

In reality, neither the model $\widehat{\mathcal{M}}_y$ nor any of the models $\widehat{\boldsymbol{\mathcal{M}}}_y$ are unidentifiable because the values of y are not observed for the entire population. Since we cannot fit the ML models to the entire population, we fit the PML models $\widehat{\boldsymbol{\mathcal{M}}}_y$ to the sample. The PMLEs of the regression coefficients $\left(\hat{\boldsymbol{\beta}}_{pmle},\hat{\boldsymbol{\sigma}}_{pmle,},\hat{\boldsymbol{\gamma}}_{pmle}\right) \in \widehat{\mathcal{M}}_y$ are consistent estimators of the MLEs of the regression coefficients $\left(\hat{\boldsymbol{\beta}}_{mle},\hat{\boldsymbol{\sigma}}_{mle,},\hat{\boldsymbol{\gamma}}_{mle}\right)$. In order to identify the true model $\mathcal{M}_y$, we use a sample-based loss function. This function does not measure the goodness of fit of the model fitted to the sample. Instead, it is an estimate of the goodness of fit of the model fitted to the entire population. Fitting the PML models and examining the values of the sample-based

loss function is intended to approximate fitting the ML models to the entire population and measuring the model goodness of fit of the population model.

Since this estimate of the population model's goodness of fit depends on the selected sample, there is uncertainty when using the models in $\widehat{\mathcal{M}}_y$ to identify the true model $\mathcal{M}_y$. However, we are not interested in measuring this uncertainty. Instead, we rank the models based on the value of the loss function and select the model with the smallest value (e.g., the best-fit model) as the sample-based estimate of the true model. In most cases, the best-fit model is the most parsimonious among the models with the lowest loss values.

1.5.5 PA Framework Definitions

DEFINITION 1.13. The PA adjustment factor is the square diagonal matrix $\hat{\mathbf{\Gamma}}_{\mathbf{X}} \in \mathbb{R}^{P\times P}$ defined as

$$\hat{\mathbf{\Gamma}}_{\mathbf{X}} = \mathbf{D}_{\mathbf{X}} \mathbf{D}_{\hat{\mathbf{X}}_w}^{-1}, \tag{1.19}$$

where $\mathbf{D}_{\mathbf{X}} = \text{diag}(\mathbf{X}) \in \mathbb{R}^{P\times P}$ is a diagonal matrix where the function $\text{diag}: \mathbb{R}^P \to \mathbb{R}^{P\times P}$ is defined as $\text{diag}(\mathbf{X}) = \sum_{k\in P} \varepsilon_k^{\mathrm{T}} \mathbf{X} \varepsilon_k \varepsilon_k^{\mathrm{T}}$ and $\varepsilon_k \in \mathbb{R}^P$ is the k-basis vector of $\mathbb{R}^P$ for $k \in \{1,...,P\}$ and $\mathbf{X} \in \mathbb{R}^P$ is a row vector $\mathbf{X} = (X_1,\ldots,X_P)$. The function $\text{diag}(\mathbf{X})$ transforms the vector $\mathbf{X}$ into a squared matrix in $\mathbb{R}^{P\times P}$ in which

the elements outside of the main diagonal are zero, and the elements on the main diagonal are the elements the vector $\mathbf{X} = (X_1, \ldots, X_P)$ as

$$\operatorname{diag}(\mathbf{X}) = \begin{bmatrix} X_1 & 0 & \ldots & 0 & 0 \\ 0 & X_2 & \ldots & 0 & 0 \\ \ldots & \ldots & \ddots & \ldots & \ldots \\ 0 & 0 & \ldots & X_{P-1} & 0 \\ 0 & 0 & \ldots & 0 & X_P \end{bmatrix}.$$

$\hat{\mathbf{X}}_w = (\mathbf{w} \odot \mathbf{x})^{\mathrm{T}} \mathbf{S} = \sum_{k \in U} w_k \mathbf{x}_k S_k$ is the vector of the HT estimators of $\mathbf{x}$ using the weights $\mathbf{w} = [w_k] \in \mathbb{R}^{N \times 1}$ (these may be the sampling weights $d_k = \pi_k^{-1}$) and $\mathbf{D}_{\hat{\mathbf{X}}_w} = \operatorname{diag}(\hat{\mathbf{X}}_w)$ is the diagonal matrix with the elements of the main diagonal being the elements of the vector $\hat{\mathbf{X}}_w$.

The large sample properties of the PA adjustment factor $\hat{\mathbf{\Gamma}}_{\mathbf{X}}$ are given by the following theorem.

THEOREM 1.1. Assume a sequence of finite populations $\{\mathcal{F}_N\}_{N=1}^{\infty}$ of increasing size $U_N = \{1, \ldots, N_N\}_{N=1}^{\infty}$ and samples $\{n_N\}_{N=1}^{\infty}$ drawn according to a sample design $\{p_N(A_N = a_N)\}_{N=1}^{\infty}$ satisfying the regularity conditions in Section 5.9 on page 252. Then the sequence of PA adjustment factors $\{\hat{\mathbf{\Gamma}}_{\mathbf{X},N}\}_{N=1}^{\infty}$ converges to the identity matrix $\mathbf{I} \in \mathbb{R}^{P \times P}$ as

$$\lim_{\substack{N\to\infty \\ n\to\infty}} \mathbb{E}\left(\hat{\mathbf{\Gamma}}_{\mathbf{X},N} - \mathbf{I} \mid \mathcal{F}\right) = \mathbf{0}. \tag{1.20}$$

See proof in Section A.3.1 in Appendix A on page 292.

DEFINITION 1.14. The PA adjusted regression coefficients $\hat{\boldsymbol{\theta}}_{pa} = \left(\hat{\boldsymbol{\beta}}_{pa}, \hat{\boldsymbol{\sigma}}_{pa}, \hat{\boldsymbol{\gamma}}_{pa}\right)^{\mathrm{T}}$. The adjustment factor $\hat{\mathbf{\Gamma}}_{\mathbf{X}}$ incorporates the population totals into the PMLEs of the regression coefficients $\hat{\boldsymbol{\theta}}_{pmle} = \left(\hat{\boldsymbol{\beta}}_{pmle}, \hat{\boldsymbol{\sigma}}_{pmle}, \hat{\boldsymbol{\gamma}}_{pmle}\right)^{\mathrm{T}}$. Let $\hat{\boldsymbol{\theta}}_{pa} = \left(\hat{\boldsymbol{\beta}}_{pa}, \hat{\boldsymbol{\sigma}}_{pa}, \hat{\boldsymbol{\gamma}}_{pa}\right)^{\mathrm{T}}$ be the PA adjusted PMLEs of the regression coefficients of the parameters of the working model $\widehat{\mathcal{M}}_y$ computed as

$$\hat{\boldsymbol{\theta}}_{pa} = \hat{\mathbf{\Gamma}}_{\mathbf{X}_\theta} \hat{\boldsymbol{\theta}}_{pmle}, \tag{1.21}$$

for $\hat{\boldsymbol{\theta}}_{pmle} \in \left\{\hat{\boldsymbol{\beta}}_{pmle}, \hat{\boldsymbol{\sigma}}_{pmle}, \hat{\boldsymbol{\gamma}}_{pmle}\right\}$, where the subscripts $\theta \in \{\beta, \sigma, \gamma\}$ of $\hat{\mathbf{\Gamma}}_{\mathbf{X}_\theta}$ indicate different subsets of auxiliary variables in the PA adjustment for the location, scale, and shape parameters. Note that the model $\widehat{\mathcal{M}}_{pa,y}$ with the adjusted parameters $\hat{\boldsymbol{\theta}}_{pa} = \left(\hat{\boldsymbol{\beta}}_{pa}, \hat{\boldsymbol{\sigma}}_{pa}, \hat{\boldsymbol{\gamma}}_{pa}\right)^{\mathrm{T}}$ is a different model from $\widehat{\mathcal{M}}_y$, except for the case when the estimated totals of the auxiliary variables match exactly to their corresponding population total for each parameter of the distribution. In this case, $\hat{\boldsymbol{\theta}}_{pa} = \hat{\boldsymbol{\theta}}_{pmle}$ because $\hat{\mathbf{\Gamma}}_{\mathbf{X}_\theta} = \mathbf{I}$. The large sample properties of $\hat{\boldsymbol{\theta}}_{pa} = \left(\hat{\boldsymbol{\beta}}_{pa}, \hat{\boldsymbol{\sigma}}_{pa}, \hat{\boldsymbol{\gamma}}_{pa}\right)^{\mathrm{T}}$ are given in the next theorem.

THEOREM 1.2. Assume a sequence of finite populations $\{\mathcal{F}_N\}_{N=1}^{\infty}$ of increasing size $U_N = \{1,...,N_N\}_{N=1}^{\infty}$ and samples $\{n_N\}_{N=1}^{\infty}$ drawn according to a sample design $\{p_N(A_N = a_N)\}_{N=1}^{\infty}$ satisfying the regularity conditions in Section 5.9 on page 252. The sequence of PA adjusted parameters $\{\hat{\boldsymbol{\theta}}_{pa,N}\}_{N=1}^{\infty}$ with $\hat{\boldsymbol{\theta}}_{pa,N} = \left(\hat{\boldsymbol{\beta}}_{pa}, \hat{\boldsymbol{\sigma}}_{pa}, \hat{\boldsymbol{\gamma}}_{pa}\right)^{\mathrm{T}}$ is design-consistent for the MLE parameters $\hat{\boldsymbol{\theta}}_{mle,N}$ in the sense that $\hat{\boldsymbol{\theta}}_{pa,N} - \hat{\boldsymbol{\theta}}_{mle,N} \mid \mathcal{F} = \mathcal{O}_p\left(n_N^{-1/2}\right)$. This result implies that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{pa,N} - \hat{\boldsymbol{\beta}}_{mle,N} \mid \mathcal{F} &= \mathcal{O}_p\left(n_N^{-1/2}\right), \\ \hat{\boldsymbol{\sigma}}_{pa,N} - \hat{\boldsymbol{\sigma}}_{mle,N} \mid \mathcal{F} &= \mathcal{O}_p\left(n_N^{-1/2}\right), \text{ and} \\ \hat{\boldsymbol{\gamma}}_{pa,N} - \hat{\boldsymbol{\gamma}}_{mle,N} \mid \mathcal{F} &= \mathcal{O}_p\left(n_N^{-1/2}\right). \end{aligned} \tag{1.22}$$

The proof is in Section A.3.3 in Appendix A. Note that the sequence of the PA adjusted parameters $\hat{\boldsymbol{\theta}}_{pa,N}$ converges in probability to the MLEs of parameters $\hat{\boldsymbol{\theta}}_{mle,N}$ of the model fitted to the N -th population in the sequence.

DEFINITION 1.15. The fitted mean $\hat{\mu}_{pa,k}$ under the PA model $\mathcal{M}_{y,pa}$. In the PA framework, we are only interested in $\hat{\mu}_{pa,k}$, the estimate of $\mu_{mle,k} = \mathbb{E}(y_k \mid \mathcal{F})$, computed by plugging the PA estimators $\left(\hat{\boldsymbol{\beta}}_{pa}, \hat{\boldsymbol{\sigma}}_{pa}, \hat{\boldsymbol{\gamma}}_{pa}\right)^{\mathrm{T}}$ in the appropriate expression of μ_k depending of the assumed model. The large sample properties of $\hat{\mu}_{pa,k}$ are given in the next theorem.

THEOREM 1.3. Assume a sequence of finite populations $\{\mathcal{F}_N\}_{N=1}^{\infty}$ of increasing size $U_N = \{1,...,N_N\}_{N=1}^{\infty}$ and samples $\{n_N\}_{N=1}^{\infty}$ drawn according to a sample design $\{p_N(A_N = a_N)\}_{N=1}^{\infty}$ satisfying the regularity conditions in Section 5.9 on page 252. The sequence of PA fitted means $\{\hat{\mu}_{pa,k,N}\}_{N=1}^{\infty}$ is design consistent for the MLE of the mean $\hat{\mu}_{mle,k,N}$ in the sense that

$$\lim_{N\to\infty} \Pr\left[\left|\hat{\mu}_{pa,k,N} - \hat{\mu}_{mle,k,N}\right| > \varepsilon_N\right] = 0, \tag{1.23}$$

for every ε_N. Note that the sequence of PA estimators $\hat{\mu}_{pa,k,N}$ converges in probability to the MLE estimator of the mean $\hat{\mu}_{mle,k,N}$ fitted to the N-th population in the sequence.

1.5.6 Miscellaneous PA Framework Definitions

DEFINITION 1.16. The canonical form of an estimator $\hat{T}$ of a population parameter T is the function f of $\boldsymbol{\pi}$ as

$$\hat{T} = \mathrm{f}(\boldsymbol{\pi}).$$

The canonical form is independent of the sample design. For example, the canonical form of the HT estimator for the total is

$$\hat{Y}_{HT} = \boldsymbol{\pi}^{\odot -1}(\mathbf{y} \odot \mathbf{S}) = \sum_{k\in A} \frac{y_k}{\pi_k},$$

where $\mathrm{f}(\boldsymbol{\pi}) = \left(\boldsymbol{\pi}^{\odot -1}\right)^{\mathrm{T}} (\mathbf{y} \odot \mathbf{S})$. The HT estimator for a SRS design is $\hat{Y}_{HT} = \frac{N}{n} \sum_{k \in A} y_k$. The canonical form of the HJ estimator (Hájek J. , 1971) is

$$\hat{Y}_{HJ} = N \frac{\left(\boldsymbol{\pi}^{\odot -1}\right)^{\mathrm{T}} (\mathbf{y} \odot \mathbf{S})}{\left(\boldsymbol{\pi}^{\odot -1}\right)^{\mathrm{T}} \mathbf{S}} = N \frac{\sum_{k \in A} \frac{y_k}{\pi_k}}{\sum_{k \in A} \frac{1}{\pi_k}},$$

where $\mathrm{f}(\boldsymbol{\pi}) = N \frac{\left(\boldsymbol{\pi}^{\odot -1}\right)^{\mathrm{T}} (\mathbf{y} \odot \mathbf{S})}{\left(\boldsymbol{\pi}^{\odot -1}\right)^{\mathrm{T}} \mathbf{S}}$. Notice that although the HJ and the HT estimators have different canonical forms, the estimators are identical for a SRS design. The canonical forms of the HT and HJ estimators of the mean are $\hat{\bar{Y}}_{HT} = \frac{\boldsymbol{\pi}^{\odot -1} (\mathbf{y} \odot \mathbf{S})}{N}$ and $\hat{\bar{Y}}_{HJ} = \frac{\left(\boldsymbol{\pi}^{\odot -1}\right)^{\mathrm{T}} (\mathbf{y} \odot \mathbf{S})}{\left(\boldsymbol{\pi}^{\odot -1}\right)^{\mathrm{T}} \mathbf{S}}$, respectively. Although the canonical forms of the HT and HJ estimators of the population mean $\bar{Y}$ are different, the estimators have the same expression in SRS designs. Note that this does not necessarily hold for other designs.

DEFINITION 1.17. There are different types of model misspecification (Rao, 1971), and we are only interested in two types. The first is when the working model has the incorrect functional form of the distribution of y. For example, let $y \in \{0,1\}$ be the outcome with a Bernoulli distribution but the distribution of the working model

is a normal distribution, and the predictions $\hat{\mu}_k$ of this model may take values different from zero or one. The second type of model misspecification includes omitted and extraneous auxiliary variables. These model misspecifications have a different impact on the efficiency of the estimators. The misspecification does not affect the consistency of the estimator because all model-assisted estimators are asymptotically unbiased and design consistent (Särndal, Swensson, & Wretman, 1992).

DEFINITION 1.18. Oracle estimator is the estimator where the functional form and auxiliary variables of the working model are not misspecified.

DEFINITION 1.19. Assuming that the same working model is fitted in the population and the sample, valid PA models are those that meet the following conditions. Both the sum of population ML residuals $\mathbf{E} = \mathbf{y} - \hat{\boldsymbol{\mu}}_{mle}$ and the weighted sum of the sample-based PML residuals $\hat{\mathbf{E}} = \mathbf{y} - \hat{\boldsymbol{\mu}}_{pmle}$ are asymptotically zero, that is

$$\mathbb{E}\left(\mathbf{1}^{\mathrm{T}}\mathbf{E}\right) = \mathcal{O}\left(\frac{1}{N}\right), \text{ and}$$
$$\mathbb{E}\left(\mathbf{1}^{\mathrm{T}}\left(\mathbf{d} \odot \mathbf{S} \odot \hat{\mathbf{E}}\right)\right) = \mathcal{O}\left(\frac{1}{n}\right).$$

This definition includes models where the sum of the residual in ML models and the weighted sum of the residuals in PML models is zero. To ensure that the models are valid, we require the intercept term to be kept in the linear regressions of all parameters of the model.

DEFINITION 1.20. A PA estimator with a working model with the vector of auxiliary variables $\mathbf{x} = (x_1, \ldots, x_P)$ and population control totals $\mathbf{X}$ is incomplete if at least one population total of an auxiliary variable x_p is estimated as $\hat{X}_{HT,p}$ rather than being known. The PA adjustment for such auxiliary variable is $\hat{\Gamma}_{X_p} = \frac{\hat{X}_{HT,p}}{\hat{X}_{HT,p}} = 1$.

DEFINITION 1.21. We describe the principles to assist estimation with full response (adapted from the principles to assist estimation in the presence of nonresponse by Särndal & Lundström, 2005). Although the PA framework can create models using many variables, it is advisable to reduce the number of candidate auxiliary variables in the collection of models by selecting variables that

i) explain the main study variable y, and

ii) explain the inclusion probabilities π if the sampling design is informative for y.

If PA estimates by domain are needed, then the auxiliary variables should also

iii) identify as closely as possible the most important domains.

Implementing the principles for estimation may require the help of subject matter experts who can determine the initial set of auxiliary variables since the PA algorithm identifies those variables that meet both conditions (i) and (ii). Implementing (iii) requires either forcing these variables in the collection of models even if they do not

explain y and π, or including the domain related variables in the selected model at the end of the PA algorithm.

REMARK 1.7. The PA framework uses matrix notation, matrix algebra, and matrix calculus to express the form of the estimators and derive their variances and estimates of variances. Dol, Steerneman, & Wansbeek (1996) show the convenience of matrix-algebra for proving the asymptotic properties of the HT estimator. Our notation emphasizes the random nature of the vector $\mathbf{S}$ that follows a discrete multinomial distribution (Tillé, 2006). The estimators and their variances are functions of $\mathbf{S}$, and are treated as random variables in multivariate statistical analysis.

1.6 Computing Algorithmic PA Estimators

As an algorithmic framework, the algorithm is the core of the production of PA estimators. The PA algorithm identifies the relevant variables that explain the outcome, taking into account the variables that explain the sample selection.

When producing the PA estimator, the algorithm incorporates the population totals of auxiliary variables into PLL of the data for an assumed working model. This information is currently ignored in the regular PML approach (Binder, 1983). The algorithm consists of 10 steps that are listed in Algorithm 1.1.

Algorithm 1.1 Algorithm for the derivation of the PA estimator

Algorithmic PA estimators	
(A) Model identification	1: Propose the collection of working models $\mathcal{M}_{\pi}$ for the inclusion probabilities π_k. 2: Identify the ML model $\widehat{\mathcal{M}}_{\pi} \in \mathcal{M}_{\pi}$ of π that minimizes the loss function L_{π}. 3: Propose the collection of working models $\mathcal{M}_{y}$ for the outcome variable y. 4: Identify the PML model $\widehat{\mathcal{M}}_{y} \in \mathcal{M}_{y}$ of y that minimizes the loss function L_{y}.
(B) Targeting of relevant variables	5: Identify the model $\mathcal{M}_{\pi,y}$ with the set of auxiliary variables that explain both y and π as $\mathcal{M}_{\pi,y} = \widehat{\mathcal{M}}_{y} \cap \widehat{\mathcal{M}}_{\pi}$. 6: Fit the PML model $\widehat{\mathcal{M}}_{\pi,y} \in \mathcal{M}_{\pi,y}$ for π using the auxiliary variables in $\mathcal{M}_{\pi,y}$ identified in Step (5). Use $\widehat{\mathcal{M}}_{\pi y}$ to compute the fitted values $\hat{\pi}_k$ and the adjusted weights $\hat{w}_k = d_k \hat{d}_k \left(\sum_{k\in U} d_k\right) / \left(\sum_{k\in U} d_k \hat{d}_k\right)$. 7: Identify the PML model $\widehat{\mathcal{M}}^{*}_{y,\hat{w}}$ of y among all models $\mathcal{M}_{y}$ that minimizes the loss function L_{y} using the adjusted weights $\hat{w}_k$ computed in Step (6).
(C) Creation of the estimator and inference	8: Fit the PML model $\widehat{\mathcal{M}}^{*}_{y}$ of y using the variables of the model $\widehat{\mathcal{M}}^{*}_{y,\hat{w}}$ identified in Step (7) using the sampling weight d_k. 9: Create the PA model $\widehat{\mathcal{M}}_{pa,y}$ by adjusting the PMLE of the regression coefficients of $\widehat{\mathcal{M}}^{*}_{y}$ from Step (8) by the PA adjustment $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$. 10: Estimate the adjusted PA fitted mean $\hat{\mu}_{pa,k}$ for $k \in A$ using the PA model $\widehat{\mathcal{M}}_{pa,y}$ from Step (9) and substitute $\hat{\mu}_{pa,k}$ in the generic form of the PA estimator $\hat{Y}_{pa} = \sum_{k\in A} d_k \hat{\mu}_{pa,k}$. Make inferences for $\hat{Y}_{PA}$ using $\hat{\mathbb{V}}\left(\hat{Y}_{PA}\right)$.

The steps were explained in detail through the example in Section 1.3 for estimates of a total of a non-negative continuous outcome and a proportion for a binary outcome. In this section, we provide additional information on computing algorithmic estimators, such as the types of outcomes and distributions of working models, alternatives for modeling the sample membership indicators such as modeling the probabilities of inclusion directly when only the sample is available, and the mathematical definition of the loss function used in the algorithm.

The algorithm is specially designed for informative sample designs, a feature not addressed by previous approaches such as Nascimiento Silva & Skinner (1997) and McConville, Breidt, Lee, & Moisen (2017). For noninformative designs, like SRS we would expect the targeted relevant variables in Steps 5 to 7 to be null, and we could skip directly to Step 8 for these designs. However, we recommend going through all steps even with noninformative designs because any particular sample outcome may be unbalanced. Going through all steps protects against unusual sample outcomes.

1.6.1 General Considerations before Computing Algorithmic PA Estimators

Before executing the algorithm, we first define the target outcome variable y and the characteristic to estimate such as a population total or population mean. The PA framework permits all types of outcomes (e.g., categorical, ordinal, continuous) and distributions of working models, although current software may limit their

computation for some distributions and variable types. The outcome variable can be a single quantity or a vector with multiple outcomes.

The complexity of the models evaluated in the PA algorithm is a function of the parameters of the working model. Although very complex models can be fitted, large samples may be needed for the PA estimator to be well-behaved (e.g., converge to a normal distribution). Since the regularity conditions for design consistency of the PA estimator require that $n_0 \gg P$, where n_0 is the observed sample size and P is the number of estimated parameters of the working model, working models with a large number of parameters relative to the sample size are not recommended. We advise following common sense rules for model building such as excluding highly correlated variables (e.g., auxiliary variables that lie entirely within the column space of $\mathbf{X}$) and variables that do not explain the outcome (e.g., the component of the candidate auxiliary variable lying outside the column space of $\mathbf{X}$ is orthogonal to $\mathbf{y}$).

The standard error of the PA estimator is estimated using the variance formulas based on Taylor series linearization (see Section 1.7). However, other methods such as replication can be used. An important element of the PA framework is the development of methods that account for the model uncertainty in the estimate of variance. Specifically, the methods should account for the effect on the variance when the models have many parameters.

REMARK 1.8. In Step 2 of Algorithm 1.1, we assume a functional form of the collection of models $\mathcal{M}_{\mathbf{y}}$ for the outcome variable y. In situations where more

than one functional form is feasible, e.g., $\mathcal{M}_{\mathbf{y}}$ and $\mathcal{M}'_{\mathbf{y}}$, the PA algorithm can be modified to select not only the auxiliary variables of the working model but also the functional form that best fits the observed sample. The AIC and dAIC, which are used to compare the goodness of fit among models, are based on likelihood/pseudo-likelihood that can accommodate models with different distributions. However, special care is needed when comparing the AIC for these models because some software packages compute the AIC ignoring the constant terms of likelihood. The difference between the AIC values of two models with the same functional form is not affected when the constant term is excluded. However, if the likelihoods of different functional forms have different constants, then the selection of the functional form is likely to be incorrect.

EXAMPLE 1.7. Returning to the estimators from Example 1.1 on page 7, algorithmic PA estimates of both the total of Y_1 (total hospital expenditures in 1998) and the proportion $\bar{Y}_2$ (proportion of hospitals that received financing from the state agency) that are likely to be efficient for both Y_1 and $\bar{Y}_2$ can be produced by identifying the common predictors of the model for both outcome variables. For example, we assume a bivariate working model $\mathcal{M}_{\mathbf{y}}$ with

$$\mathbf{y}_k \sim \mathbf{N}\left(\begin{pmatrix}\eta_{\beta_1 k}\\ \eta_{\beta_{12} k}\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}^2\\ \sigma_{12}^2 & \sigma_2^2\end{pmatrix}\right),$$

where the outcome vector is $\mathbf{y}_k = (y_{1k}, y_{2k})^{\mathrm{T}}$, $\eta_{\beta_1 k} = \mathbf{x}_{\beta_1 k}\boldsymbol{\beta}_1$ is the linear predictor associated with y_1, $\eta_{\beta_2 k} = \mathbf{x}_{\beta_2 k}\boldsymbol{\beta}_2$ is the linear predictor associated with y_2, $\mathbf{x}_{\beta_1 k}$ is the vector of auxiliary variables associated with y_1, and $\mathbf{x}_{\beta_2 k}$ is the vector of auxiliary variables associated with y_2. Note that the working models in the collection of $\mathcal{M}_{\mathbf{y}}$ are misspecified because the support of the variable y_1 (total hospital expenditures) is $y_{1k} \geq 0$ while $\hat{y}_{1k} \in \mathbb{R}$, and the support of y_2 (indicator whether or not the hospital received state agency funds) is $\{0,1\}$ while $\hat{y}_2 \in \mathbb{R}$. Since we want to identify common variables that explain both $\mathbf{y}_k$ and S_k, we recommend using the same vector of auxiliary variables for $\mathbf{x}_{\beta_2 k}$, $\mathbf{x}_{\beta_2 k}$, and $\mathbf{x}_{\pi k}$ when defining the collections $\mathcal{M}_{\mathbf{y}}$ and $\mathcal{M}_{\pi}$. If there are no common variables among the models except for the intercept term, then the PA estimator is the poststratified estimator to the total population size. The models are fitted using multivariate regression subroutines or by fitting the models for the outcome separately.

EXAMPLE 1.8. In Example 1.1 on page 7, the PA adjusted fitted means $\hat{\mu}_{pa,k}$ for y_1 (hospital expenditures in 1998) can be negative because the assumed working model is normal with a linear location parameter. The negative values may be an issue for totals of some small domains. We discuss two ways to ensure that $\hat{\mu}_{pa,k}$ for y_1 is always nonnegative (assuming that the regularity conditions for the MLE estimators for $\hat{\mu}_{pa,k}$ hold). The first is to use the same linear model but with a

different link function; for example, $\log\left(y_{1k}\right)=\mathbf{x}_k\boldsymbol{\beta}$ so $\hat{\mu}_{pa,y_1,k}=\exp\left(\mathbf{x}_k\hat{\boldsymbol{\beta}}_{pa}\right)\geq 0$ for $k\in A$. The second is to assume a different working model with an appropriate support; for example, the exponential distribution $y\overset{iid}{\sim}\mathcal{E}xp\left(\theta_\beta\right)$ with a probability density function $\mathrm{f}_Y\left(y;\theta_\beta\right)=\theta_\beta\exp\left(-\theta_\beta y\right)1_{\{y\geq 0\}}$ where $\theta_\beta=\mathbf{x}_k\boldsymbol{\beta}$. A similar approach is used to ensure $\hat{\mu}_{pa,k}\in\left(0,1\right)$ for the binary variable y_2. For example, we can assume a working model $y_{2k}\sim\mathcal{B}e\left(\theta_\beta\right)$ where $\mathrm{logit}\left(\theta_{\beta k}\right)=\mathbf{x}_k\boldsymbol{\beta}$. The previous three working models yield nonlinear algorithmic PA estimators (see Definition 1.23 on page 90). However, even though linear and nonlinear estimators converge to the same limit for working models with the same number of auxiliary variables, the MSE of a nonlinear estimator is larger than the MSE of a linear estimator with the same size when the sample size is small. In other words, when the sample sizes are small, the sample size of a nonlinear estimator needed to achieve the same MSE of a linear estimator is larger than the sample size of the linear estimator. The difference in MSE is also a function of the sample design and the complexity of the distribution of the working model. The differences in efficiency between linear and nonlinear PA estimators are empirically studied in Section 2.2.

1.6.2 Alternative Models for S

As mentioned in Definition 1.2, there are different ways to model $\mathbf{S}$ depending on the availability of the frame and probability of inclusions for the PA models. We identify four situations:

A. When the sample selection indicator S_k for $k \in U$ is modeled using the complete population or frame,

B. When the inclusion probability π_k for $k \in U$, instead of the sample selection indicator, is modeled directly using the complete frame,

C. When the sample selection indicator S_k for $k \in A$ is modeled using the sample, and

D. When the inclusion probability π_k for $k \in A$, instead of the sample selection indicator, is modeled directly using the sample.

Algorithm 1.1 creates the algorithmic PA estimator for situation A and is described in detail in Example 1.1. In this example, the sample membership indicator S_k for $k \in U$ is the dependent variable with a collection of models $\mathcal{M}_\pi$ with an assumed working model $s_k \mid \mathcal{F} \overset{iid}{\sim} \mathcal{B}e(\pi_k)$ where $\pi_k = \text{logit}^{-1}(\mathbf{x}_k\boldsymbol{\beta})$ that are fitted using ML since the frame is available. In situation B, π_k is fitted, instead of S_k, assuming a different working model since π_k is a continuous variable in a range $\pi_k \in (0,1)$. One

possible model for π_k is the fractional logit with $\pi_k = \text{logit}^{-1}(\mathbf{x}_k\boldsymbol{\beta})$ (See Remark 1.3 for a discussion of alternative models for π_k). The working models of S_k or π_k in situations A and B are fitted using ML using the frame. In situations C and D, the working models of S_k or π_k are fitted using PML using the sample (see Remark 1.6).

EXAMPLE 1.9 We illustrate the impact on the precision of the algorithmic PA estimators under situations A through D using alternative working models for S_k or π_k (Bernoulli, fractional logistic, and linear models) fitted to either the population or sample using the sample design and population from Example 1.1.

Table 1.7 shows the empirical relative efficiency (RE) of nine algorithmic PA estimators of Y_1 and nine estimators of $\bar{Y}_2$ compared with the HT estimator using 100,000 draws (see the definition of the RE in Section A.4 in Appendix A). The algorithmic PA estimators are identified by the number in the rows named "Estimator #" on the table. The last column of the table shows the RE of the GREG VDK estimators for the same population characteristics. The table shows that all algorithmic PA estimators fitted to either the population or the sample using MLE or PMLE are more efficient than the HT estimators of Y_1 and $\bar{Y}_2$. The algorithmic PA estimators are also more efficient than the HJ estimators, which are not included in the table.

Table 1.7 shows that the algorithmic PA estimators for Y_1 and $\overline{Y}_2$ (estimators 2 and 6, respectively), with a assumed fractional logistic working model fitted to the frame using ML, are slightly more efficient than the PA estimators 1 and 5 with an assumed Bernoulli working model also fitted to the frame using ML. When the model is fitted to the observed sample, the algorithmic PA estimators of Y_1 and $\overline{Y}_2$ (estimators 4 and 8), with aa assumed fractional logistic working model fitted using PML, are slightly more efficient than the PA estimators 3 and 7 with an assumed Bernoulli working model fitted using PML fitted to the sample.

Although the differences are very small, all algorithmic PA estimators with assumed Bernoulli or fractional logistic working models fitted to either the frame or sample (estimators 1 to 8) are more efficient than the VDK estimators despite the uncertainty in identifying the model. The minimum and maximum RE differences between the PA estimators and VDK estimators are 0.51 and 0.64 percentage points for Y_1, and 2.43 and 3.26 percentage points for $\overline{Y}_2$. The largest differences correspond to the PA estimators with the fractional logistic model for π_k fitted to the sample (PA estimators 3 and 7).

When the assumed working model of π_k is the linear probability model (estimators 9 through 16, see Remark 1.3), the algorithmic PA estimators are slightly more efficient than the VDK estimators except for the PA estimators of Y_1 with the linear models $S_k = \mathbf{x}_k\boldsymbol{\beta}$ and $\pi_k = \mathbf{x}_k\boldsymbol{\beta}$ fitted to the sample (estimators 9 and 10). In contrast, the same PA estimators of Y_1 fitted to the frame are more efficient than the VDK

estimators (11 and 12) with somewhat larger differences in RE. For the estimators of $\bar{Y}_2$ with a linear probability model, the maximum and minimum differences in RE between the PA estimators 13 to 16 and the VDK are generally less than one percentage point.

Table 1.7 Relative efficiency compared to HT of the algorithmic PA estimators and VDK by alternative models for estimating $\hat{\mathcal{M}}_\pi$ in Example 1.1

Method	Estimator				
	Algorithmic PA				VDK
	MLE		PMLE		GREG
Data file	Population	Population	Sample	Sample	Sample
Dependent variable	S_k	π_k	S_k	π_k	N/A
Situation	A	B	C	D	N/A
Model $\hat{\mathcal{M}}_\pi$	Bernoulli	Fractional logistic	Bernoulli	Fractional logistic	N/A
Relative efficiency (HT)					
Estimator #	(1)	(2)	(3)	(4)	
Total Y_1	7.56	7.63	7.63	7.68	7.04
Estimator	(5)	(6)	(7)	(8)	
Proportion $\bar{Y}_2$	77.76	77.87	77.19	78.02	74.76
Model $\hat{\mathcal{M}}_\pi$	Linear	Linear	Linear	Linear	N/A
Relative efficiency (HT)					
Estimator #	(9)	(10)	(11)	(12)	
Total Y_1	6.19	6.72	8.24	7.68	7.03
Estimator #	(13)	(14)	(15)	(16)	
Proportion $\bar{Y}_2$	77.19	77.71	77.66	78.02	74.76

Although no generalizations are possible based on the results of one simulation study, the gains in efficiency may be larger if we assume a more complex working model that matches the type of data for π_k. However, these gains may be very small as

illustrated in this example. These results also suggest that modeling π_k instead S_k may yield more efficient algorithmic PA estimators independently of fitting the model to the frame or the sample. One reason may be that $S_k = s_k$ is a dichotomized version of π_k, which generally leads to a loss of information (Kotsiantis & Kanellopoulos, 2006). Since the goal of the PA algorithm is to identify the relevant auxiliary variables that explain the sample selection, modeling S_k may add unnecessary noise. Although there are no differences in RE between the algorithmic PA estimators fitted to the frame with assumed Bernoulli and fractional logistic working models, we hypothesize these models are practically the same because of the large frame.

1.6.3 The Loss Function

In the PA algorithm, the comparisons among the fitted models in $\boldsymbol{\mathcal{M}}$ in Steps 2, 4, and 7 use a loss function, $\mathrm{L}(\mathcal{M}): \mathbb{R} \to \mathbb{R}$, that measures the goodness of fit of the models $\mathcal{M} \in \boldsymbol{\mathcal{M}}$ being evaluated. In the PA algorithm, when the model is fitted using ML, the loss function is based on the Akaike information criterion (AIC), see Akaike (1981). The AIC is an estimator of the quality of a model relative to others for a given set of data. The AIC is used for variable selection in model building (Hastie, Tibshirani, & Friedman, 2009). The AIC is computed as

$$\mathrm{AIC}\left(\widehat{\mathcal{M}}\right) = 2P - 2\mathcal{L}\left(\widehat{\mathcal{M}}\right), \tag{1.24}$$

where P is the number of parameters fitted in the model $\widehat{\mathcal{M}}$ and $\mathcal{L}\left(\widehat{\mathcal{M}}\right)$ is the maximum value of the likelihood of the fitted model $\widehat{\mathcal{M}}$. Smaller values of the AIC indicate better goodness of fit. The first term in (1.24) penalizes the AIC by the number of estimated parameters to prevent overfitting.

In the PA algorithm, when the model is fitted to the observed sample using PML, the loss function is a design-based version of the AIC defined as

$$\text{dAIC}\left(\widehat{\mathcal{M}}\right) = 2P - 2\mathcal{L}\left(\widehat{\mathcal{M}} \mid \mathcal{F}\right),$$

where $\mathcal{L}\left(\widehat{\mathcal{M}} \mid \mathcal{F}\right)$ is the maximum value of the PL of the fitted model $\widehat{\mathcal{M}}$. The $\text{dAIC}\left(\widehat{\mathcal{M}}\right)$ is an estimate of the $\text{AIC}\left(\widehat{\mathcal{M}}\right)$, that is, the AIC of the model $\mathcal{M}$ fitted to the entire population. The loss function for the model $\mathcal{M}_y$ fitted to the population is $L\left(\widehat{\mathcal{M}}\right) = \text{AIC}\left(\widehat{\mathcal{M}}\right)$ and for the model fitted to the sample is $L\left(\widehat{\mathcal{M}} \mid \mathcal{F}\right) = \text{dAIC}\left(\widehat{\mathcal{M}} \mid \mathcal{F}\right)$.

Although $\boldsymbol{\mathcal{M}}_\pi$ and $\boldsymbol{\mathcal{M}}_y$ are collections of infinite number of working models, the PA algorithm does not fit all models nor evaluate their loss functions. Instead, a subset of candidate models is generated using a one-variable-at-a-time stepwise forward variable selection based on the value of the AIC or dAIC depending on whether the model is fitted to the sample or frame. This method of variable selection is a greedy algorithm that adds the best variable and removes the worst one from the working model at each step measuring the goodness of fit on the AIC/dAIC for each

variable addition and deletion. The algorithm attempts to find a global optimum through optimal local decisions in each step (Guyon & Elisseeff, 2003; Tang, Alelyani, & Liu, 2014). This approach reduces the algorithm computation time because not all models are fitted and evaluated.

The appeal of the AIC is the simplicity of the expression that does not require multiple statistical tests of the coefficients of the linear estimators of the model parameters[5]. Since the AIC is a relative measure among working models, the selected model may have a poor fit if none of the models describes the observed data well. In the PA approach, the poor fit of the working models is not a major issue because the resulting algorithmic PA estimator, as any model-assisted estimator, is always design-consistent even if the working model is misspecified.

REMARK 1.9. The stepwise AIC variable is a commonly used method for model building (Rawlings, Pantula, & Dickey, 1998); however, there are criticisms since some of its assumptions are violated when used in this way. These criticisms are important for standard statistics but are not necessarily a weakness within the PA framework. These criticisms of the AIC are most relevant when the prediction of

[5] When the observations are *iid* for linear regression, the one-variable-at-a-time AIC stepwise selection is asymptotically equivalent to the stepwise selection using a cut-off for *p*-values of about 15.7 percent. This is equivalent to comparing two models using the likelihood ratio test (Heinze, Wallisch, & Dunkler, 2018). This relationship has not been shown for the sample-based AIC.

future observations is the goal[6]. The models selected using the AIC may suffer from selection bias since the variables with a large explanatory power in the observed sample are more likely to be selected (Heinze, Wallisch, & Dunkler, 2018). The selected model may not be the best to predict future samples. In contrast, this property is desirable in the PA framework because the PA estimator is derived from the observed sample and used to adjust the same observed sample and not for adjustments of future samples. In other words, we are interested in the variables that have large explanatory power.

REMARK 1.10. Although we have chosen the AIC as the loss function for the PA algorithm, any other sample-based metric for measuring the goodness of fit such as the adjusted R^2 and Schwarz or Bayesian information criterion (BIC) can be used provided that there is a theoretical justification for the sample design and the availability of software that computes these metrics (see Section A.2 in Appendix A on page 290 for the theoretical justification of the sample-based AIC, dAIC, used in the PA algorithm). Among the methods for variable selection, we do not recommend those that rely on hypothesis testing such as stepwise regression based on *p*-values, *F*-tests, *t*-tests of the regression coefficients or model fit statistics. The reliability of the modified tests that reflect the sample design requires relatively large samples

[6] Prediction in this context is the process for determining the value of statistical variables at some future point in time. This type of prediction is not relevant within the survey-sampling context. This prediction is also not to be confused with the model-based estimation methodology from Valliant, Dorfman, & Royal (2000) where predictions refer to as the values of cases not selected in the sample.

(Mukhopadhyay, 2016). We also do not recommend variable selection methods that rely on regularization because their goal is to minimize MSE instead of the bias (Hastie, Tibshirani, & Friedman, 2009)[7]. Generally, these methods do not reflect the effect of the sample design in the variable selection. Although the LASSO can be used as a method for variable selection for complex designs (McConville, Breidt, Lee, & Moisen, 2017), our empirical results show that when the model is not sparse, LASSO tends to select fewer variables in the working model. Selecting fewer variables is the opposite of the goal of the PA algorithm; that is, identifying all relevant variables related to the outcome of the working model (see discussion in Section 4.6).

1.6.4 Implementation of the PA Algorithm and Computation of PA Estimators

The PA estimators, algorithm, and evaluation in this article are implemented in R (R Development Core Team, 2017) with modifications under the GNU General Public License (GPL-2, https://www.r-project.org/Licenses/GPL-2) to the R packages *sampling* (Tillé & Matei, 2016), *survey* (Lumley, 2012), *GAMLSS* (Rigby & Stasinopoulos, 2005), and the core statistics of R (R Development Core Team, 2017).

[7] Although a large variance may be problem, the primary goal of the model selection is to reduce the bias. Once this has achieved, methods to reduce the variance can be used when the variance is large.

1.7 Statistical Properties of the Algorithmic PA Estimator

In this section, we present the generic expressions of the PA estimator, variance, and variance estimator. We also derive the large sample or asymptotic properties of the PA estimators using the approach from Fuller (2009) and Isaki & Fuller (1982), which is the standard for studying the large-sample properties of estimators in survey sampling theory. In this setting, we assume an indexed sequence of nested finite populations $\{\mathcal{F}_N\}_{N=1}^{\infty}$ of size N_N and the associated sequence of sample designs $p(A_N = a_N)$ that meet suitable regularity conditions listed in Section 5.7. We show that the sequence of PA estimators $\left\{\hat{Y}_{PA,N}\right\}_{N=1}^{\infty}$ is design consistent for the finite population total Y_N in the N-th population with a limiting normal distribution that allows inferences about the finite population total through tests of hypothesis or confidence intervals.

1.7.1 The Generic Form of the PA Estimator and its Design-Based Asymptotic Properties

Although the specific form of the PA estimator is only known at the end of the algorithm, we can study the properties of a generic form of the algorithmic PA estimator. Assume a superpopulation model $\mathcal{M}_y$ for the outcome variable y, a finite population $\mathcal{F}$ consisting of N *iid* realizations from the superpopulation that is sampled according to sample design $p(A = a)$ as described in Section 1.5.3. We are

interested in estimating the population total $Y = \sum_{k \in U} y_k$, or the population mean $\bar{Y} = \frac{Y}{N}$ based on the realized sample $A = a$. The generic expression of the PA estimator of the population total Y, $\hat{Y}_{PA}$, is

$$\hat{Y}_{PA} = \mathbf{w}^{\mathrm{T}} \left(\hat{\boldsymbol{\mu}}_{pa} \odot \mathbf{S} \right) = \sum_{k \in U} w_k \hat{\mu}_{pa,k} S_k \,, \tag{1.25}$$

where $\mathbf{w} = \left[w_k \right] \in \mathbb{R}^{N \times 1}$ is the vector of the weights described in Section 1.7.4, and $\hat{\boldsymbol{\mu}}_{pa} = \left[\hat{\mu}_{k,pa} \right] \in \mathbb{R}^{N \times 1}$ is the vector of the PA adjusted fitted means $\hat{\mu}_{pa,k}$ of the working model computed as $\hat{\mu}_{pa,k} = \mathbb{E}\left(\mathbf{g}^{-1}\left(\mathbf{x}_k \hat{\boldsymbol{\beta}}_{pa} \right) \right)$.

The following results describe the asymptotic properties of the generic PA estimator.

THEOREM 1.4. Assume a sequence of finite populations $\left\{ \mathcal{F}_N \right\}_{N=1}^{\infty}$ of increasing size $U_N = \left\{ 1, ..., N_N \right\}_{N=1}^{\infty}$ and samples $\left\{ n_N \right\}_{N=1}^{\infty}$ drawn according to a sample design $\left\{ p_N \left(A_N = a_N \right) \right\}_{N=1}^{\infty}$ satisfying the regularity conditions listed in Section 5.9. Let $\left\{ \hat{Y}_{PA,N} \right\}_{N=1}^{\infty}$ be the sequence of PA estimators $\hat{Y}_{PA,N}$ of the total Y_N in the N-th population. Then $\left\{ \hat{Y}_{PA,N} \right\}_{N=1}^{\infty}$ is design consistent of the population total Y_N in the sense that

$$\frac{\hat{Y}_{pa,N}}{N} - \frac{Y_N}{N} \middle| \mathcal{F}_N = \mathcal{O}_p\left(n_N^{-1/2}\right). \tag{1.26}$$

The immediate result of Theorem 1.4 is that the variance of the sequence of PA estimators $\left\{\hat{Y}_{PA,N}\right\}_{N=1}^{\infty}$ is stochastically bounded in the sense that

$$\mathbb{V}\left(\frac{\hat{Y}_{pa,N}}{N_N} - \frac{Y_N}{N_N} \middle| \mathcal{F}_N \right) = \mathcal{O}\left(n_N^{-1}\right). \tag{1.27}$$

The limiting distribution of the sequence of PA estimators $\left\{\hat{Y}_{PA,N}\right\}_{N=1}^{\infty}$ is

$$\frac{\hat{Y}_{PA,N} - Y_N}{\sqrt{\mathbb{V}\left(\hat{Y}_{PA,N} \mid \mathcal{F}_N\right)}} \xrightarrow{D} \mathcal{N}(0,1), \tag{1.28}$$

where $\mathcal{N}(0,1)$ is the standard normal distribution. Similarly, the limiting distribution of the sequence of PA estimators $\left\{\hat{Y}_{PA,N}\right\}_{N=1}^{\infty}$ when $\mathbb{V}\left(\hat{Y}_{PA,N} \mid \mathcal{F}_N\right)$ is estimated by $\hat{\mathbb{V}}\left(\hat{Y}_{PA,N} \mid \mathcal{F}_N\right)$ is

$$\frac{\hat{Y}_{PA,N} - Y_N}{\sqrt{\hat{\mathbb{V}}\left(\hat{Y}_{PA,N} \mid \mathcal{F}_N\right)}} \xrightarrow{D} \mathcal{N}(0,1). \tag{1.29}$$

The proofs of these results are found in Section 5.9.

1.7.2 Specific Forms of the PA Estimator and their Expressions of Variance

In general, the estimator $\hat{Y}_{PA}$ is nonlinear, so we approximate $\mathbb{V}\left(\hat{Y}_{PA}\right)$ using the linear terms of the Taylor's Series (TS) expansion of $\hat{Y}_{PA} = Z(\mathbf{S})$. Let $Z : \mathbb{R}^N \to \mathbb{R}$ be a vector-to-scalar valued function of $\mathbf{S}$ where $Z(\mathbf{S}) = \left(\mathbf{w} \odot \hat{\boldsymbol{\mu}}_{pa}\right)^{\mathrm{T}} \mathbf{S}$. The function $Z(\mathbf{S})$ is approximated by the linear terms of the multivariate TS expansion evaluated at point $\mathbb{E}(\mathbf{S}) = \boldsymbol{\pi}$ (see Section 5.9 in Chapter 4). Then the approximate variance of $\hat{Y}_{PA}$ is

$$\mathrm{A}\,\mathbb{V}\left(\hat{Y}_{PA}\right) = N^2 \mathbf{Z}'(\boldsymbol{\pi})^{\mathrm{T}}\, \boldsymbol{\Delta}\, \mathbf{Z}'(\boldsymbol{\pi}), \tag{1.30}$$

where $\mathbf{Z}'(\boldsymbol{\pi}) = \dfrac{\partial \mathbf{Z}(\mathbf{S})}{\partial \mathbf{S}} \mathbf{Z}(\boldsymbol{\pi}) = \left.\dfrac{\partial \mathbf{Z}(\mathbf{S})}{\partial \mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}}$ (with some abuse of notation) is the vector of the directional partial derivatives of $\mathbf{Z}$ with respect to $\mathbf{S}$, evaluated at $\mathbf{S} = \boldsymbol{\pi}$, and $\boldsymbol{\Delta}$ is the variance-covariance matrix of $\mathbf{S}$. The approximate variance of $\hat{Y}_{PA}$, $\mathbb{AV}\left(\hat{Y}_{PA}\right)$, can be interpreted as the variance of the HT estimator of the linear substitutes $z_k \in \mathbf{Z}'(\boldsymbol{\pi})$ for $k \in N$ (Woodruff, 1971).

The variance estimator of $\hat{Y}_{PA}$ is

$$\hat{\mathbb{V}}\left(\hat{Y}_{PA}\right) = N^2 \hat{\mathbf{Z}}'^{\mathrm{T}}(\boldsymbol{\pi}) \hat{\boldsymbol{\Delta}}\, \hat{\mathbf{Z}}'(\boldsymbol{\pi}), \tag{1.31}$$

where $\hat{\mathbf{Z}}'(\boldsymbol{\pi})$ is the partial derivatives with respect to $\mathbf{Z}$ after replacing the unknown quantities by their sample-based estimates, $\hat{\Delta} = \Delta \oslash \left(\Delta + \boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}}\right)$, and $\oslash$ is the Hadamard division operator.

REMARK 1.11. The algebraic expression of the partial derivatives in the vector $\mathbf{z}$ can be difficult to derive for some nonlinear PA estimators, specifically those PA estimators that use calibrated weights, because the weights are also functions of $\mathbf{S}$. One approach is to numerically compute the partial derivatives $z_k \in \mathbf{Z}'(\boldsymbol{\pi})$ for $k \in A$ and substitute the numeric vector in (1.31), following an approach similar to Woodruff & Causey (1976)[8]. Although the algebraic expressions of the partial derivatives are not needed since they are numerically computed; this approach still requires the functional form of $\mathbf{Z}(\mathbf{S})$. Another alternative is to use replication methods to estimate the variance $\mathbb{V}\left(\hat{Y}_{PA}\right)$. See Section 5.9.4 for computing the variance and variance estimator for a nonlinear PA estimator with a Poisson distribution and the log link function.

REMARK 1.12. Demnati & Rao (2004) and Shah (2004) comment on the issue with the TS linearization method for survey sampling estimates, which can produce different variance estimators that are all asymptotically design-unbiased. They argue that choice of the appropriate variance estimator requires considering an

[8] Higher-order methods for numerical approximation of the partial derivative are available in some R packages.

assumed model and the validity of that model under repeated sampling. Demnati & Rao (2004) developed a TS linearization approach for deriving variance estimators that leads directly to a unique expression of the variance based on smooth functions of totals. In the PA approach, the expressions of the estimates of variance are also unique and match those expressions from Demnati & Rao (2004). The difference is that the variances in the PA approach are based on functions of the random vector $\mathbf{S}$. Since the estimators are linear/nonlinear functions of random variables, the expressions of the variances are computed using the methods for computing the variances of functions of the random variable $\mathbf{S}$.

EXAMPLE 1.10. Table 1.8 shows the expressions of PA estimators for totals and variances estimator for some working models. The estimators of the means are obtained by dividing the estimators of the total by N and the variance by N^2.

Based on Definition 1.19, the sum of the residuals at the population level defined as $\mathbf{E}_{mle} = \mathbf{y} - \hat{\boldsymbol{\mu}}_{mle}$ is asymptotically zero, and the weighted sum of the residuals at the sample level defined as $\hat{\mathbf{E}}_{pmle} = \left(\mathbf{y} - \hat{\boldsymbol{\mu}}_{pmle}\right) \odot \mathbf{S}$ is also asymptotically zero in valid PA working models. However, there is a second type of residuals defined as $\mathbf{e}_{mle} = \mathrm{g}^{-1}\left(\mathbf{y}_k\right) - \mathbf{x}_k \hat{\boldsymbol{\beta}}_{mle}$ for the population and $\breve{\mathbf{e}}_{pmle} = \left(\mathrm{g}^{-1}\left(\mathbf{y}_k\right) - \mathbf{x}_k \hat{\boldsymbol{\beta}}_{pmle}\right) \odot \mathbf{S}$ for the sample. The sum of the residuals $\mathbf{e}_{mle}$ in the population and the weighted sum of

Table 1.8 PA estimators of the total Y and their variance estimators

Estimator	Point estimator	Variance estimators
Horvitz-Thompson (HT)	$\hat{Y}_{HT} = \mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{s})$	$\hat{\mathbb{V}}\left(\hat{Y}_{HT}\right) = (\mathbf{y} \odot \mathbf{d} \odot \mathbf{s})^{\mathrm{T}} \hat{\boldsymbol{\Delta}} (\mathbf{y} \odot \mathbf{d} \odot \mathbf{s})$ where $\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$
Hájek (HJ)	$\hat{Y}_{HJ} = N \dfrac{\mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{s})}{\mathbf{d}^{\mathrm{T}}\mathbf{s}}$	$\hat{\mathbb{V}}\left(\hat{Y}_{HJ}\right) = \dfrac{N^2}{\hat{N}_{HT}^2} (\breve{\mathbf{e}} \odot \mathbf{d} \odot \mathbf{s})^{\mathrm{T}} \hat{\boldsymbol{\Delta}} (\breve{\mathbf{e}} \odot \mathbf{d} \odot \mathbf{s})$ where $\breve{\mathbf{e}} = \left(\mathbf{y} - \hat{\bar{Y}}_{HJ}\right) \odot \mathbf{s}$, $\hat{N}_{HT} = \mathbf{d}^{\mathrm{T}}\mathbf{s}$
Ratio (RA)	$\hat{Y}_{RA} = X \dfrac{\mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{s})}{\mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{s})}$	$\hat{\mathbb{V}}\left(\hat{Y}_{RA}\right) = \left(\dfrac{X}{\hat{X}_{HT}}\right)^2 (\breve{\mathbf{e}} \odot \mathbf{d} \odot \mathbf{s})^{\mathrm{T}} \hat{\boldsymbol{\Delta}} (\breve{\mathbf{e}} \odot \mathbf{d} \odot \mathbf{s})$ where $\breve{\mathbf{e}} = \left(\mathbf{y} - \mathbf{x}\hat{R}_{HT}\right) \odot \mathbf{s}$, , $\hat{R}_{HT} = \dfrac{\hat{Y}_{HT}}{\hat{X}_{HT}}$, $\hat{X}_{HT} = \mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{s}), \hat{Y}_{HT} = \mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{s})$
Normal Distribution with identity link function (GREG)*	$\hat{Y}_{Normal} = \mathbf{X}\hat{\boldsymbol{\beta}}_{mle}$ where $\hat{\boldsymbol{\beta}}_{mle} = \hat{\mathbf{T}}_{\mathbf{xx}}^{-1} \hat{\mathbf{T}}_{\mathbf{xy}}$ $\hat{\mathbf{T}}_{\mathbf{xx}} = (\mathbf{x} \odot \mathbf{s})^{\mathrm{T}} (\mathbf{d} \odot \mathbf{x} \odot \mathbf{s})$, $\hat{\mathbf{T}}_{\mathbf{xy}} = (\mathbf{x} \odot \mathbf{s})^{\mathrm{T}} (\mathbf{d} \odot \mathbf{y} \odot \mathbf{s})$	$\hat{\mathbb{V}}\left(\hat{Y}_{normal}\right) = \mathbf{X}\hat{\mathbf{T}}_{\mathbf{x},\mathbf{x}}^{-1} (\mathbf{x} \odot \mathbf{d} \odot \breve{\mathbf{e}} \odot \mathbf{s})^{\mathrm{T}} \hat{\Delta} (\mathbf{x} \odot \mathbf{d} \odot \breve{\mathbf{e}} \odot \mathbf{s}) \hat{\mathbf{T}}_{\mathbf{xx}}^{-1} \mathbf{X}^{\mathrm{T}}$ where $\breve{\mathbf{e}} = \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right) \odot \mathbf{s}$
Poisson distribution with log link function	$\hat{Y}_{Poisson} = \mathbf{d}^{\mathrm{T}}\left(\hat{\boldsymbol{\mu}}_{pa} \odot \mathbf{s}\right)$ where $\hat{\boldsymbol{\mu}}_{pa} = \exp\left((\mathbf{s} \odot \mathbf{x})\hat{\boldsymbol{\beta}}_{pa}\right)$	See Section 5.9.4.

*See derivation of the PA estimator in Section A.5 on Appendix A on page 304.

residuals $\breve{\mathbf{e}}_{pmle}$ in the sample are also asymptotically zero. The two types of residuals, $\mathbf{E}_{mle}$ and $\mathbf{e}_{mle}$ for the population and $\mathbf{E}_{pmle}$ and $\mathbf{e}_{pmle}$ are exactly zero when the link function is the identity function. The importance of the second type of residual is its use in computing the variance as illustrated in the following remark.

REMARK 1.13 Another expression for the variance of model-assisted estimator is based on the HT variance of the variable for the residuals defined as $\varepsilon_k = y_k - \mathrm{m}(y_k)$. The variance is

$$\mathbb{V}\left(\hat{Y}\right) = N^2 \sum_{k\in U} \sum_{l\in U} \left(\pi_{kl} - \pi_k \pi_l\right) \frac{\varepsilon_k}{\pi_k} \frac{\varepsilon_l}{\pi_l}, \tag{1.32}$$

where $\mathrm{m}(y_k)$ is a model-based estimator of $\mu_k = \mathbb{E}(y_k)$. Similarly, the expression of the variance estimator is

$$\hat{\mathbb{V}}\left(\hat{Y}\right) = N^2 \sum_{k\in A} \sum_{l\in A} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{\varepsilon_k}{\pi_k} \frac{\varepsilon_l}{\pi_l}. \tag{1.33}$$

The expressions (1.32) and (1.33) are derived in Wu & Sitter (2001), Breidt & Opsomer, (2017), and Breidt & Opsomer (2000). Särndal & Lundström (2005) recommend these expressions when computing the variance for the GREG estimators with residuals $\varepsilon_k = y_k - \mathbf{x}_k \hat{\boldsymbol{\beta}}_{pmle}$. These expressions are different from the variance of the PA/GREG estimator in Table 1.8. The variance estimator of the PA estimator includes the factors $\mathbf{X}\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\left(\mathbf{x}\odot\mathbf{d}\odot\breve{\mathbf{e}}\odot\mathbf{s}\right)^{\mathrm{T}}$ and $\left(\mathbf{x}\odot\mathbf{d}\odot\breve{\mathbf{e}}\odot\mathbf{s}\right)\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\mathbf{X}^{\mathrm{T}}$ where

$\breve{\mathbf{e}} = \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right) \odot \mathbf{s}$. This factor represents the g-weights used in the alternative expressions of the variance and variance estimator of the GREG estimator:

$$\mathbb{V}\left(\hat{Y}\right) = N^2 \sum_{k \in A} \sum_{l \in A} \left(\pi_{kl} - \pi_k \pi_l\right) \frac{g_k \varepsilon_k}{\pi_k} \frac{g_l \varepsilon_l}{\pi_l} \text{ and} \tag{1.34}$$

$$\hat{\mathbb{V}}\left(\hat{Y}\right) = N^2 \sum_{k \in A} \sum_{l \in A} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{g_k \varepsilon_k}{\pi_k} \frac{g_l \varepsilon_l}{\pi_l}, \tag{1.35}$$

where $g_k = 1 + \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right) \mathbf{T}_{\mathbf{xx}}^{-1} \mathbf{x}_k$ (see Särndal & Lundström, 2005). In other words, the expression of the variance estimator of the PA/GREG estimator in Table 1.8 is equal to the expression of the variance of the GREG estimator with the g-weights in (1.35). The PA approach naturally accounts for the g-weights that are more appropriate on theoretical grounds (see Särndal, Swensson, & Wretman 1989). Looking at the asymptotic properties, the g-weights converge in probability to 1 since $g_N - 1 = \mathcal{O}_p\left(n^{-1}\right)$. Thus (1.35) approaches (1.34) as $g_l \varepsilon_l \to \varepsilon_l$ in large samples. In other words, the variance and variance estimators with the g-weights are more appropriate for smaller samples since they adjust for the discrepancies between the auxiliary variable population totals $\mathbf{X}$ and the estimates of these population totals $\hat{\mathbf{X}}_{HT}$ in the observed sample.

A close examination of the variances of other PA estimators in Table 1.8 shows that they also have factors similar to the g-weights that converge in probability to 1 in large samples. Table 1.9 lists the "g-weights" factors for other estimators listed in Table 1.8.

Breidt & Opsomer (2017) and Särndal & Lundström (2005) suggest ignoring the g-weights in the variance estimator because they are asymptotically one. However, relying on asymptotic consistency may not be justified when the sample is small. Furthermore, standard practice for the other estimators such as the HJ and RA estimators does not ignore their g-weights in their estimated variances. Ignoring these g-weights in the variance estimator ignores the auxiliary variables, which is precisely the information we want to include to reduce the variance.

Table 1.9 The g-weights like factors in some PA estimators

Estimator	g-weight factor
Horvitz-Thompson (HT)	1
Hájek (HJ)	$\frac{N}{\hat{N}_{HT}}$
Ratio (RA)	$\frac{X}{\hat{X}_{HT}}$
Normal distribution with identity link function (GREG)	$\mathbf{X}\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\left(\mathbf{x}\odot\mathbf{d}\odot\breve{\mathbf{e}}\odot\mathbf{s}\right)^{\mathrm{T}}$, $\left(\mathbf{x}\odot\mathbf{d}\odot\breve{\mathbf{e}}\odot\mathbf{s}\right)\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\mathbf{X}^{\mathrm{T}}$

1.7.3 Linear and Nonlinear PA Estimators

We refer to PA estimators as linear or nonlinear depending on how the auxiliary variables are related to the outcome variable.

DEFINITION 1.22. A PA estimator[9] is linear if its working model is a fixed effect normal distribution $\mathcal{N}\left(\mu_k, \sigma^2\right)$ with an identity link function $\mu_k = \mathbb{E}\left(y_k \mid \mathbf{x}_k\right) = \mathbf{x}_k\boldsymbol{\beta}$. The generic expression (1.25) for the linear PA estimator is

$$\hat{Y}_{PA} = \mathbf{d}^{\mathrm{T}}\left(\mathbf{x}_k\hat{\boldsymbol{\beta}}_{pa} \odot \mathbf{s}\right) = \sum_{k\in U} d_k s_k \mathbf{x}_k\hat{\boldsymbol{\beta}}_{pa}, \tag{1.36}$$

where $\mathbf{d} \in \mathbb{R}^{N\times 1}$, $\mathbf{d} = \left[d_k\right]$, and d_k is the sampling weight, $\hat{\boldsymbol{\beta}}_{pa} \in \mathbb{R}^{P\times 1}$ are the PA adjusted PMLE regression coefficients computed as $\hat{\boldsymbol{\beta}}_{pa} = \hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\hat{\boldsymbol{\beta}}_{pmle}$, where $\hat{\boldsymbol{\beta}}_{pmle} \in \hat{\mathcal{M}}_{plme,y}$ are the PMLEs of $\boldsymbol{\beta} \in \mathcal{M}_y$, $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}} \in \mathbb{R}^{P\times P}$ is the PA adjustment (see Definitions 1.12 and 1.13), and $\mathbf{s} = \left[s_k\right] \in \left\{0,1\right\}^{N\times 1}$, where s_k is the realized sample membership indicator for $k \in U$.

We implicitly refer to a linear PA estimator or linear working model when the working model meets Definition 1.22 unless stated otherwise. Cassel, Särndal, & Wretman's (1977) definition for linear estimators in survey sampling theory is $\hat{\theta} = \beta_{s_0} + \sum_{k\in A}\beta_{sk}y_k$ and focuses on the linear combinations of the outcome variable instead of the parameters and auxiliary variables of the model.

THEOREM 1.5. The linear PA estimators can be written as the weighted sum of the population totals of the auxiliary variables of the PA working model $\hat{\mathcal{M}}_{PA,y}$ as

---

[9] This classification is similar to the linear and nonlinear GREG estimators in Särndal (2007)).

$$\hat{Y}_{PA} = \mathbf{X}\,\hat{\boldsymbol{\beta}}_{pmle}. \tag{1.37}$$

The proof follows after replacing $\hat{\boldsymbol{\beta}}_{pa}$ by $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\hat{\boldsymbol{\beta}}_{pmle}$ in (1.25) using the sample design weights d_k for $k \in U$.

Theorem 1.5 shows that if the working model is linear, the estimate of the total Y is a function of the PML estimates of regression coefficients $\boldsymbol{\beta}$ of the working model. One immediate result of this theorem is the following corollary:

COROLLARY 1.1. The variance of the linear PA estimator is

$$\mathbb{V}\left(\hat{Y}_{PA} \mid \mathcal{F}\right) = \mathbf{X}^{\mathrm{T}}\mathbb{V}\left(\hat{\boldsymbol{\beta}}_{pmle} \mid \mathcal{F}\right)\mathbf{X}, \tag{1.38}$$

which is a function of the variance of the parameters of the working model. Although this expression looks like a model-based estimator, it is a design-based estimator, and its variance depends on the sample design.

REMARK 1.14 The expression (1.37) is the form of the linear generalized regression (GREG) estimator (Särndal, Swensson, & Wretman, 1992) with as assisting model with $\mathbb{E}(y_k) = \mathbf{x}_k\boldsymbol{\beta}$ and $\mathbb{V}(y_k) = \sigma^2$. The $\hat{Y}_{PA}$, computed as the sample weighted sum of the PA adjusted PMLE means of a normal model or as the sum of products of the PMLE of $\boldsymbol{\beta}$ and their associated population totals, reproduces the GREG estimator $\hat{Y}_{GREG} = \hat{Y}_{HT} + \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right)\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = \hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}}$, $\hat{\mathbf{T}}_{\mathbf{xx}} = \mathbf{x}^{\mathrm{T}}\left(\mathbf{d} \odot \mathbf{x} \odot \mathbf{S}\right)$, and $\hat{\mathbf{T}}_{\mathbf{xy}} = \mathbf{x}^{\mathrm{T}}\left(\mathbf{d} \odot \mathbf{y} \odot \mathbf{S}\right)$. However, the PA linear estimator

and the linear GREG estimator are not the same since the set of auxiliary variables in the working model of the PA estimator is random that depends on the sample, while the auxiliary variables in the linear GREG estimator are fixed. The linear GREG estimators are a subclass of the PA linear estimator. These results are not surprising since both are extremum estimators that optimize mathematically equivalent criterion functions when they have the same working model (Greene, 2008). Fitting a well-defined working normal model and using the PMLEs of the regression coefficients of the working model produces the same model-assisted estimator when the assisting model is used to guide the form of the estimator.

EXAMPLE 1.11. Some examples of PA linear estimators and their corresponding parameters of the normal working model are listed in Table 1.10.

Table 1.10 Examples of linear PA estimators

Estimator name	Working model	Estimator	Notes
Hájek	$y_k \overset{iid}{\sim} \mathcal{N}\left(\beta, \sigma^2\right)$	$\hat{Y} = N\hat{\beta}_{pmle}$	$\hat{\beta}_{pmle} = \frac{\sum_{k \in A} d_k y_k}{\sum_{k \in A} d_k}$
Stratified	$y_k \overset{iid}{\sim} \mathcal{N}\left(\beta_h, \sigma_h^2\right)$, for $h \in \{1, ..., H\}$	$\hat{Y} = \sum_{h=1}^{H} N_h \hat{\beta}_{pmle,h}$	$\hat{\beta}_{pmle,h} = \frac{\sum_{k \in A_h} d_k y_k}{\sum_{k \in A_h} d_k}$
Ratio	$y_k \overset{iid}{\sim} \mathcal{N}\left(\beta x_k, \sigma^2 x_k\right)$	$\hat{Y} = X\hat{\beta}_{pmle}$	$\hat{\beta}_{pmle} = \frac{\sum_{k \in A} d_k y_k}{\sum_{k \in A} d_k x_k}$
Linear regression one variable x_k	$y_k \overset{iid}{\sim} \mathcal{N}\left(\beta x_k, \sigma^2\right)$	$\hat{Y} = X\hat{\beta}_{pmle}$	$\hat{\beta}_{pmle} = \frac{\sum_{k \in A} d_k x_k y_k}{\sum_{k \in A} d_k x_k^2}$
Stratified separate ratio	$y_k \overset{iid}{\sim} \mathcal{N}\left(\beta_h x_k, \sigma_h^2 x_k\right)$, for $h \in \{1, ..., H\}$	$\hat{Y} = \sum_{h=1}^{H} X_h \hat{\beta}_{pmle,h}$	$\hat{\beta}_{pmle,h} = \frac{\sum_{k \in A_h} d_k y_k}{\sum_{k \in A_h} d_k x_k}$
Stratified combined ratio	$y_k \overset{iid}{\sim} \mathcal{N}\left(\beta x_k, \sigma^2 x_k\right)$, for $h \in \{1, ..., H\}$	$\hat{Y} = \sum_{h=1}^{H} X_h \hat{\beta}_{pmle}$	$\hat{\beta}_{pmle} = \frac{\sum_{h=1}^{H} \sum_{k \in A_h} d_k y_k}{\sum_{h=1}^{H} \sum_{k \in A_h} d_k x_k}$

DEFINITION 1.23. All PA estimators that do not meet Definition 1.22 are called nonlinear estimators. The nonlinear PA estimators are new and differ from the nonlinear GREG estimators described in Särndal (2007) and Breidt & Opsomer (2017). Closed form expressions of nonlinear estimators often do not exist, and they must be computed numerically. The expression of the nonlinear estimator depends on the distribution of the working model. For example, if the working model is a non-normal generalized linear model (GLM), the nonlinear PA estimator is

$$\hat{Y}_{PA} = \sum_{k \in A} w_k \, \mathrm{g}^{-1}\left(\mathbf{x}_k \hat{\boldsymbol{\beta}}_{pa,k}\right), \tag{1.39}$$

where $\hat{\boldsymbol{\beta}}_{pa} \in \mathbb{R}^{P \times 1}$ are the PA adjusted PML estimates of the PMLE of the regression coefficients $\hat{\boldsymbol{\beta}}_{pmle}$ computed as $\hat{\boldsymbol{\beta}}_{pa} = \hat{\boldsymbol{\Gamma}}_{\mathbf{X}} \hat{\boldsymbol{\beta}}_{pmle}$, and g^{-1} is the inverse of the link function. The PMLEs of the coefficients of the linear predictor, $\hat{\boldsymbol{\beta}}_{pmle}$, are computed maximizing the PL using iteratively reweighted least squares (IRLS) in combination with numerical algorithms such as Gauss-Newton and Levenberg–Marquardt. Nonlinear PA estimators can always be computed when the auxiliary variable population totals are available; in contrast, nonlinear GREG estimators require complete auxiliary information (i.e., all $\mathbf{x}_k$ are known).

EXAMPLE 1.12. In Section 2.2 on page 130, we evaluate the performance of three nonlinear PA estimators with assumed working models based on Bernoulli, Poisson, and Gamma distributions. Table 1.11 lists the working models and functional forms of the nonlinear PA estimators from Section 2.2. The table also

includes other nonlinear PA estimators with different with other nonlinear working models.

REMARK 1.15 Särndal (2007) defines nonlinear GREG estimators as those that are generated by working models other than linear fixed effects models. Although this definition almost matches Definition 1.23 for nonlinear PA estimators, there are important differences. The nonlinear GREG estimator is based on two working models: a nonlinear primary model used to derive an auxiliary variable and population total and a linear secondary working model that is to produce the functional form the estimator. To illustrate the role of the primary and secondary working models, assume we want to compute a nonlinear GREG estimator using a GLM model for the variable y_k with $\mathbb{E}(y_k \mid \mathbf{x}_k) = \mathrm{g}^{-1}(\mathbf{x}_k\boldsymbol{\beta})$. Since y is only observed in the sample, a primary PL nonlinear model $\widehat{\mathcal{M}}_{pmle,y}$ with the auxiliary variables $\mathbf{x}$ is fitted and used to compute the PMLEs of the regression coefficients $\hat{\boldsymbol{\beta}}_{pmle}$. The same model $\widehat{\mathcal{M}}_{pmle,y}$ is then used to predict the estimated means $\hat{\mu}_{pmle,k} = \mathrm{g}^{-1}\left(\mathbf{x}_k\hat{\boldsymbol{\beta}}_{plme}\right)$ for all elements of the population. Note that this requires knowing all the values of $\mathbf{x}_k$ for $k \in U$. The fitted PL mean $\hat{\mu}_{pmle,k}$ of the primary

Table 1.11 Examples of nonlinear PA estimators

Nonlinear PA Estimator	Working model y_k	Link function $\mathrm{g}(\theta)$	Expression
Bernoulli	$\mathcal{B}e(\theta_k)$	$\mathbf{x}_k\boldsymbol{\beta} = \mathrm{logit}(\theta_k)$	$\hat{Y} = \sum_{k\in A} d_k \dfrac{e^{\mathbf{x}_k\hat{\boldsymbol{\beta}}_{pa}}}{1+e^{\mathbf{x}_k\hat{\boldsymbol{\beta}}_{pa}}}$
Poisson	$\mathcal{P}o(\theta_k)$	$\mathbf{x}_k\boldsymbol{\beta} = \log(\theta_k)$	$\hat{Y} = \sum_{k\in A} d_k e^{\mathbf{x}_k\hat{\boldsymbol{\beta}}_{pa}}$
Gamma	$\mathcal{G}(\theta_k)$	$\mathbf{x}_k\boldsymbol{\beta} = \log(\theta_k)$	$\hat{Y} = \sum_{k\in A} d_k e^{\mathbf{x}_k\hat{\boldsymbol{\beta}}_{pa}}$
Lognormal	$\mathrm{Log}\mathcal{N}\left(\theta_{\beta,k},\theta_{\sigma k}^2\right)$	$\mathbf{x}_k\boldsymbol{\beta} = \theta_{\beta k}$, $\mathbf{x}_{\sigma,k}\boldsymbol{\sigma} = \log\left(\theta_{\sigma,k}\right)$	$\hat{Y} = \sum_{k\in A} d_k \exp\left(\mathbf{x}_{\beta,k}\hat{\boldsymbol{\beta}}_{pa} + \exp\left(\mathbf{x}_{\sigma,k}\boldsymbol{\sigma}\right)^2/2\right)$
Inverse Gaussian	$\mathcal{IG}\left(\theta_{\beta,k},\theta_{\sigma k}\right)$	$\mathbf{x}_k\boldsymbol{\beta} = \dfrac{1}{\theta_{\beta k}^2}$,	$\hat{Y} = -\sum_{k\in A} \dfrac{d_k}{\mathbf{x}_k\hat{\boldsymbol{\beta}}_{pa}}$
Exponential	$\mathcal{E}xp\left(\theta_{\beta,k}\right)$	$\mathbf{x}_k\boldsymbol{\beta} = -\dfrac{1}{\theta_{\beta k}}$	$\hat{Y} = \sum_{k\in A} d_k \sqrt{\mathbf{x}_k\hat{\boldsymbol{\beta}}_{pa}}$
Zero-inflated Poisson	$\left(\theta_{\alpha,k} + \left(1-\theta_{\alpha,k}\right)e^{-\theta_{\beta,k}}\right)1_{\{y_k=0\}}$ $+\left(\left(1-\theta_{\alpha,k}\right)\dfrac{\theta_{\beta,k}^{y_k}e^{-\theta_{\beta,k}}}{y_k!}\right)1_{\{y_k\in\mathbb{N}_{>0}\}}$	$\mathbf{x}_{\alpha,k}\boldsymbol{\alpha} = \log\left(\theta_{\alpha,k}\right)$ $\mathbf{x}_{\beta k}\boldsymbol{\beta} = \mathrm{logit}\left(\theta_{\beta,k}\right)$	$\hat{Y} = \sum_{k\in A} d_k \left(\dfrac{e^{\mathbf{x}_{\beta,k}\hat{\boldsymbol{\beta}}_{pa}}}{1+e^{\mathbf{x}_{\beta,k}\hat{\boldsymbol{\beta}}_{pa}}}\right) e^{\mathbf{x}_{\alpha,k}\hat{\boldsymbol{\alpha}}_{pa}}$

model and the estimated population total $\widetilde{M} = \sum_{k \in U} \hat{\mu}_{pmle,k}$ are estimates of the ML mean $\hat{\mu}_{mle,k}$ of the working model fitted to the population and the sum of the means $M = \sum_{k \in U} \hat{\mu}_{mle,k}$, respectively. The tilde (~) indicates that the population total $\widetilde{M}$ is not computed as the HT estimator of the fitted means as $\hat{M}_{HT} = \sum_{k \in A} d_k \hat{\mu}_{pmle,k}$, but rather as the sum of the predictions $\hat{\mu}_{pmle,k}$ for each element in the population. Since the population total $\widetilde{M}$ is an estimate of M, then the variance of the estimated total $\widetilde{M}$ is $\mathbb{V}\left(\widetilde{M} \mid \mathcal{F}\right) \neq 0$ because the value of $\widetilde{M}$ depends on the selected sample. At this step, the auxiliary variables $\mathbf{x}$ from the primary model are discarded, and the derived auxiliary variable $\hat{\mu}_{pmle}$ and population total $\widetilde{M}$ are used in a secondary normal working model to form a linear GREG estimator. The secondary working model is $\mathcal{N}\left(\mathbf{m}_k \boldsymbol{\alpha}, \sigma^2\right)$, with location parameters $\boldsymbol{\alpha} = \left(\alpha_0, \alpha_1\right)^{\mathrm{T}}$. The auxiliary variables are $\mathbf{m}_k = \left(1, \hat{\mu}_{pmle,k}\right)$, and population totals are $\widetilde{\mathbf{M}} = \left(N, \widetilde{M}_k\right)$. The general expression of the nonlinear GREG estimator of the total Y is the linear estimator

$$\hat{Y}_{NLGREG} = \hat{Y}_{HT} + \left(\widetilde{\mathbf{M}} - \hat{\mathbf{M}}_{HT}\right)\hat{\boldsymbol{\alpha}}, \tag{1.40}$$

where $\hat{\boldsymbol{\alpha}} = \left(\hat{\alpha}_0, \hat{\alpha}_1\right)^{\mathrm{T}} \in \mathbb{R}^{2\times 1}$ are the linear regression estimators of $\boldsymbol{\alpha}$ computed as $\hat{\boldsymbol{\alpha}} = \hat{\mathbf{T}}_{\mathbf{mm}}^{-1}\hat{\mathbf{T}}_{\mathbf{my}}$, where $\hat{\mathbf{T}}_{\mathbf{mm}} = \sum_{k \in A} d_k \mathbf{m}_k^{\mathrm{T}} \mathbf{m}_k$, $\hat{\mathbf{T}}_{\mathbf{my}} = \sum_{k \in A} d_k \mathbf{m}_k^{\mathrm{T}} y_k$, $\hat{\mathbf{M}}_{HT} = \left(\hat{N}_{HT}, \hat{M}_{HT}\right)$, and $\hat{N}_{HT} = \sum_{k \in A} d_k$.

From the PA context, the nonlinear GREG estimators are incomplete PA estimators (see Definition 1.20) with a derived variable (See Section 1.8) and a normal model $\mathcal{N}\left(\mathbf{m}_k\boldsymbol{\alpha},\sigma^2\right)$. Note that if the model is correct, we expect that $\hat{\alpha}_0 = 0$ and $\hat{\alpha}_1 = 1$.

One of the earliest nonlinear GREG estimators described in the literature is the logistic generalized regression estimator (LGRE) from Lehtonen & Veijanen (1998). In the simple case, the LGRE estimator assumes that the primary working model of the outcome y_k is $y_k|\mathbf{x}_k \overset{iid}{\sim} \mathcal{B}e\left(\theta_k\right)$ with a link function $\text{logit}\left(\theta_k\right) = \mathbf{x}_k\boldsymbol{\beta}$, and the mean $\mu_k = \mathbb{E}\left(y_k\right) = \dfrac{\exp\left(\mathbf{x}_k\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_k\boldsymbol{\beta}\right)}$ estimated using PML as $\hat{\mu}_{pmle.k}$. The estimated population total $\widetilde{M} = \sum_{k\in U}\hat{\mu}_{pmle.k}$, is the sum of the derived auxiliary variable $\hat{\mu}_{pmle,k.}$ The secondary working model is $y_k \overset{iid}{\sim} \mathcal{N}\left(\hat{\mu}_{mle,k}\alpha,\sigma^2\right)$, which is linear on the fitted PMLE mean $\hat{\mu}_{pmle.k}$ of the first model. The expression of the Lehtonen & Veijanen (1998) nonlinear GREG estimator for the total Y is (1.40) after substituting $\hat{\boldsymbol{\alpha}}$, $\widetilde{\mathbf{M}}$, and $\hat{\mathbf{M}}_{HT}$ by $\hat{\alpha} = \dfrac{1}{\hat{N}_{HT}}\sum_{k\in A} d_k\hat{\mu}_{pmle,k}$, $\widetilde{M}$, and $\hat{M}_{HT}$, respectively.

Wu & Sitter (2001) propose a nonlinear GREG estimator called a model calibrated (MC) estimator. They follow the same approach described above and produce two versions of MC estimators based on two secondary working models. The primary model is the same as described above. The primary model is fitted to the population to derive the population total as $\widetilde{M} = \sum_{k\in U}\hat{\mu}_{pmle,k}$. The secondary model

of the first MC estimator is $\mathcal{N}\left(\mathbf{m}_k\boldsymbol{\alpha},\sigma^2\right)$ described above, and the expression of the first MC estimator is (1.40). For SRS, which is the sample design used in Wu & Sitter (2001), then $\hat{N}_{HT} = N$ and (1.40) reduces to

$$\hat{Y}_{MC} = \hat{Y}_{HT} + \left(\widetilde{M} - \hat{M}_{HT}\right)\hat{\alpha}_1. \tag{1.41}$$

The secondary working model of the second version of the MC estimator is $y_k \overset{iid}{\sim} \mathcal{N}\left(\hat{\mu}_{pmle,k}\hat{\alpha}_1,\sigma^2\right)$, and the expression of the estimator of the total Y is

$$\hat{Y}_{MC} = \hat{Y}_{HT} + \left(\widetilde{M} - \hat{M}_{HT}\right)\frac{\sum_{k\in A} d_k y_k \hat{\mu}_{mle,k}}{\sum_{k\in A} d_k \hat{\mu}^2_{mle,k}}, \tag{1.42}$$

which is the calibration estimator with one auxiliary variable $\hat{\mu}_{mle,k}$ and the estimated population total $\widetilde{M}$.

All the nonlinear GREG estimators described above require the values of the auxiliary variables to be known (e.g., complete auxiliary information) for computing the estimated population total.

The properties and performance of the linear and nonlinear PA estimators compared to the linear and nonlinear GREG estimators are studied through simulation in Section 2.2. The results indicate that linear and nonlinear PA estimators have approximately the same performance as the linear and nonlinear GREG estimators when the appropriate weight w_k is used in (1.25), and the use of complete auxiliary

information does not improve the efficiency of the nonlinear GREG estimators for the evaluated models.

1.7.4 Alternative Weights for Nonlinear PA Estimators

In SRS designs, the sampling weights d_k always meet the calibration equations $\sum_{k\in A} d_k = N$ and $\sum_{k\in A} d_k\pi_k = n$ (see Kott, 2006 for the definition of calibration equations). For designs other than SRS, the nonlinear PA estimators require very large samples to converge compared to the sample size needed with the linear estimators. One way to improve the rate of convergence in PA nonlinear estimators is to replace the weights d_k by calibrated weights w_k in the PA estimator in (1.25). We have studied three options for the weight w_k. These are:

1. The sample design weights $d_k = \frac{1}{\pi_k}$;
2. The weights calibrated (e.g., poststratified) to the population size N, defined as $w_{k,(N)} = \frac{d_k N}{\sum_{k\in A} d_k}$ (i.e., $\sum_{k\in A} w_{k,(N)} = N$); and
3. The weights calibrated using raking to both the population size N, and the sample size n denoted as $w_{k,(N,n)}$ such as the calibration equations $\sum_{k\in A} w_{k,(N,n)} = N$ and $\sum_{k\in A} w_{k,(N,n)}\pi_k = n$ are met where $n = \mathbb{E}\left(\sum_{k\in A} d_k\pi_k \mid \mathcal{F}\right)$.

All these sets of weights—d_k, $w_{k,(N)}$, and $w_{k,(N,n)}$ for $k \in A$,—produce sequences of PA estimators that are asymptotically equivalent in the sense that

$$N_N^{-1}\left(\hat{Y}_{pa,N} - \hat{Y}_{w(N),N}\right) = \mathcal{O}_p\left(n_N^{-1/2}\right), \text{ and}$$
$$N_N^{-1}\left(\hat{Y}_{pa,N} - \hat{Y}_{w(N,n),N}\right) = \mathcal{O}_p\left(n_N^{-1/2}\right).$$

However, Le Cam (1986) notes that the asymptotic theory does not inform on the estimator properties for finite sample sizes found in practice. Since the estimators $\hat{Y}_{PA}$, $\hat{Y}_{w(N)}$, and $\hat{Y}_{w(N,n)}$ are asymptotically equivalent, we may just as well use any of them in large samples. Le Cam's point is demonstrated later in Section 2.2 when we find substantial differences in efficiency among nonlinear estimators for different weights and sample designs with small samples. The PA framework attempts to find consistent estimators that also have good finite sample size efficiency.

In probability proportional to size (PPS) designs, the PA estimator using the weight $w_{k,(N,n)}$ tends to be more efficient, and the gain in efficiency is greater in nonlinear working models. In Poisson (PO) sample designs, where the sampling weights d_k do not meet either the calibration equation, the weights $w_{k,(N,n)}$ can achieve large gains in efficiency for both linear and nonlinear models as shown in the examples in Section 2.2. This result justifies the practice of calibrating sampling weights as a preliminary step before additional adjustments as done in Brick, Flores Cervantes, Lee, & Norman (2011).

1.7.5 Bias-Corrected PA Estimators

According to Definition 1.18, valid PA models are those where both the sum of the maximum likelihood (ML) residuals or the weighted sum of pseudo-maximum likelihood (PML) residuals are asymptotically zero. This restriction limits somewhat the models that can be used for the creation of PA estimators. However, the expression of the PML estimator with an invalid model can be modified to ensure that the sum of the residuals is zero, at least in expectation. The resulting bias corrected PA estimator is still asymptotically unbiased and design consistent. The modification of the expression of the bias adjusted PA estimator is illustrated in the following example.

EXAMPLE 1.13. Define a collection of models $\mathcal{M}_y$ for the outcome variable y, where $y_k \mid x_k \overset{iid}{\sim} \mathcal{N}\left(\beta x_k, \sigma^2 x_k^{\gamma}\right)$ with one auxiliary variable x_k and population total X. The collection of models $\mathcal{M}_y$ defines a family of normal ratio estimators with parameters θ with a location prediction $\eta_\beta = x_k\beta$; a scale predictor $\eta_\sigma = x_k\sigma$; and shape predictor $\eta_\gamma = \gamma$ for different values of γ. We use identity link functions are used for the three parameters. Among the ratio estimators produced for $\gamma = \{0,1,2\}$ shown in Table 1.12, only those with $\gamma = \{0,1\}$ are valid PA models.

Examining the creation of the PMLE estimator for the shape parameter $\gamma = 2$ in the last row of the table, the value of $\hat{\beta}_{pmle}$ is obtained by solving the sample based

estimating equation (e.g., the estimating equation is the partial derivative of the PL function with respect to β set to zero)

$$\frac{\partial \mathcal{L}\left(\beta,\sigma^2;x,d\,|\,\mathcal{F}\right)}{\partial \beta}=\mathcal{S}\left(\beta\,|\,\mathcal{F}\right)=\sum_{k\in A} d_k \frac{\left(y_k-x_k\beta\right)}{x_k}=0, \tag{1.43}$$

and the solution is $\hat{\beta}_{pmle}=\hat{\bar{r}}_{HT}$, where $\hat{\bar{r}}=\frac{\hat{r}_{HT}}{\hat{N}_{HT}}$, $\hat{r}_{HT}=\sum_{k\in A} d_k\hat{r}_k$, $\hat{r}_k=\frac{y_k}{x_k}$ and $\hat{N}_{HT}=\sum_{k\in A} d_k$. We know that this is not a valid PA model because if the sum of the weighted residuals is zero, $\sum_{k\in A} d_k e_k=0$ where $e_k=y_k-\hat{\beta}_{pmle}x_k$, then $\hat{\beta}_{pmle}=\frac{\hat{Y}_{HT}}{\hat{X}_{HT}}\neq\hat{Y}_{HT}$. Although we cannot remove the bias completely, we can remove it in expectation by creating a difference estimator using the estimators $\hat{Y}_{PA}$ and $\hat{Y}_{PMLE}$ as

$$\hat{Y}_{PA,adj}=\hat{Y}_{HT}+\left(\hat{Y}_{PA}-\hat{Y}_{PMLE}\right), \tag{1.44}$$

where $\hat{Y}_{PMLE}$ is the estimator of the population total Y from the PML model identified in Step 8 of Algorithm 1.1 computed as $\hat{Y}_{PA}=\sum_{k\in A} d_k\hat{\mu}_{pa,k}$, and $\hat{Y}_{PA}$ is the PA estimator created in Step 9 of Algorithm 1.1 as $\hat{Y}_{PMLE}=\sum_{k\in A} d_k\hat{\mu}_{pmle,k}$. In this case, the estimator for the model for the ratio for $\gamma=2$ is

$$\hat{Y}_{PA,adj}=\hat{\bar{r}}_{HT}X+\left(\hat{Y}_{HT}+\hat{\bar{r}}_{HT}\hat{X}_{HT}\right). \tag{1.45}$$

If the total population N is known, then the PA bias adjusted estimator of the mean $\hat{\bar{Y}}$ is

$$\bar{Y}_{PA,adj} = \hat{\bar{r}}_{HT}\,\bar{X} + \left(\frac{\hat{Y}_{HT}}{N} + \hat{\bar{r}}_{HT}\,\frac{\hat{X}_{HT}}{N} \right), \tag{1.46}$$

which generalizes the Hartley-Ross ratio estimator for the mean for SRS to any sample design. The Hartley-Ross ratio estimator under SRS is $\hat{\bar{Y}}_{HR} = \bar{r}\,\bar{X} + \frac{N-1}{N}\frac{n}{n-1}\left(\bar{y} - \bar{r}\,\bar{x}\right)$, where $\frac{N}{N-1} \approx 1$ and $\frac{n}{n-1} \approx 1$ (Hartley & Ross 1954).

Table 1.12 Normal ratio models and their associated PMLE estimators for Example 1.1

Shape parameter γ	Model	PMLE Estimator $\hat{Y} = X\hat{\beta}_{pmle}$	Description	Valid PA estimator?
0	$\mathcal{N}\left(\beta x_k, \sigma_0^2\right)$	$\hat{\beta}_{pmle} \dfrac{\sum_{k\in A} d_k x_k y_k}{\sum_{k\in A} d_k x_k^2}$	GREG with one auxiliary variable x_k .	Yes
1	$\mathcal{N}\left(\beta x_k, \sigma_0^2 x_k\right)$	$\hat{\beta}_{pmle} = \dfrac{\sum_{k\in A} d_k y_k}{\sum_{k\in A} d_k x_k}$	Classical ratio estimator	Yes
2	$\mathcal{N}\left(\beta x_k, \sigma_0^2 x_k^2\right)$	$\hat{\beta}_{pmle} = \dfrac{\sum_{k\in A} d_k \frac{y_k}{x_k}}{\sum_{k\in A} d_k}$	Design biased ratio	No

The generic expression for the bias adjusted PA estimator for a total is

$$\hat{Y}_{PA,adj} = \sum_{k \in A} d_k \left(y_k + \hat{\mu}_{pa,k} - \hat{\mu}_{pmle,k} \right). \tag{1.47}$$

Note that if the weighted residuals add to zero, then the expression (1.47) becomes (1.25) with $w_k = d_k$ for $k \in A$. The variance of the bias-corrected nonlinear estimators is more difficult to obtain since it requires the linearization of $\hat{\mu}_{pa,k}$ and $\hat{\mu}_{pmle,k}$. Still, the general formula in Section 1.7.1 applies.

Table 1.13 shows the general expression of the bias-corrected normal ratio models for any value of γ. The second row shows the special case for the collection of models for a Poisson design with units sampled with probabilities of inclusion π_k.

Although the steps of Algorithm 1.1 (or Algorithm 3.1 for algebraic estimators) for creating bias-corrected PA estimators are straightforward, software to produce the estimators may not be available. For example, the value of the shape parameter γ for linear regression models can be estimated using the package `gamlss` (Stasinopoulos et al., 2017); however, the function is unstable when the location and scale parameters of the model do not include intercept terms (e.g., $\eta_\beta = \beta_1 x_k$, and $\eta_\sigma = \sigma_1 x_k$). The package `lmvar` (Posthuma Partners, 2018) is more stable, but does not fit models using PMLE nor does it does produce the AIC for the evaluation of the model. Thus, solving for the estimates would require a large programming effort.

Table 1.13 Bias-corrected PA estimators for normal ratio models

Working model	Estimator	PMLE estimator of regression coefficient $\hat{\beta}_{mle}$	Notes
$y_k \overset{iid}{\sim} \mathcal{N}\left(\beta x_k, \sigma^2 x_k^{\gamma}\right)$	$\hat{Y} = \hat{\beta}_{pmle} X + \left(\hat{Y}_{HT} - \hat{X}\right)$	$\hat{\beta}_{pmle} = \dfrac{\sum_{k \in A} d_k y_k x_k^{1-\gamma}}{\sum_{k \in A} d_k x_k^{2-\gamma}}$	For $x_k \neq c \in \mathbb{R}$
$y_k \overset{iid}{\sim} \mathcal{N}\left(\beta \pi_k, \sigma^2 \pi_k^2\right)$	$\hat{Y} = \hat{\beta}_{pmle} n + \left(\hat{Y}_{HT} - n_s\right)$	$\hat{\beta}_{pmle} \dfrac{\sum_{k \in A} d_k^2 y_k}{\sum_{k \in A} d_k}$	Poisson designs n_s is the observed sample size and n is the expected sample size.

EXAMPLE 1.14. In this example, we show the flexibility of the PA approach for producing estimators from different types of models. We assume a superpopulation multivariate model to describe the joint distribution of the study variable y and the auxiliary variables $\mathbf{x}$ that are also assumed to be random. Unlike previous examples, we do not assume a model with a univariate distribution based on the linear regression model $\mathbb{E}(y) = \boldsymbol{\beta}\mathbf{x}$.

Let $\mathbf{z}_k = (y_k, \mathbf{x}_k) \in \mathbb{R}^{1\times(P+1)}$ be one realization generated from the superpopulation model $\mathcal{M}_z$ with a multivariate normal defined by

$$\mathbf{z}_k \overset{idd}{\sim} \mathcal{N}\left(\begin{pmatrix}\beta_y \\ \boldsymbol{\beta}_{\mathbf{x}}^{\mathrm{T}}\end{pmatrix}, \begin{pmatrix}\sigma_{y_1}^2 & \boldsymbol{\Sigma}_{y\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{x}y} & \boldsymbol{\Sigma}_{\mathbf{xx}}\end{pmatrix}\right), \tag{1.48}$$

with $y_k \in \mathbb{R}$ is the study variable, $\mathbf{x}_k = (x_{1k}, ..., x_{Pk}) \in \mathbb{R}^{P\times 1}$ is the vector of the auxiliary variables, where $\boldsymbol{\beta} = \left(\beta_y, \boldsymbol{\beta}_{\mathbf{x}}^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^{P+1\times 1}$ is the vector of the location parameters of $\mathcal{M}_z$, $\beta_y \in \mathbb{R}$ is the location parameter of y_k, $\boldsymbol{\beta}_{\mathbf{x}} = (\beta_1, ..., \beta_P)^{\mathrm{T}} \in \mathbb{R}^P$ are the location parameters of $\mathbf{x}_k$, $\boldsymbol{\Sigma}_{\mathbf{xx}} = \left[\sigma_{x_p x_q}^2\right] \in \mathbb{R}^{P\times P}$ is the variance-covariance matrix of $\mathbf{x}$ where $\sigma_{x_p x_q}^2$ is the covariance between x_p and x_q for $p, q \in \{1, ..., P\}$, $\boldsymbol{\Sigma}_{y\mathbf{x}} = \left[\sigma_{yx_p}^2\right] \in \mathbb{R}^{1\times P}$ where $\sigma_{yx_p}^2$ is the variance-covariance vector between y and $\mathbf{x}$ for $p \in \{1, ..., P\}$, and $\boldsymbol{\Sigma}_{y\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}y}^{\mathrm{T}}$. Assume that the population totals $(N, \mathbf{X})$ are

known. Let $\mathcal{F}=(\mathbf{y},\mathbf{x})$ be the generated finite population as N *iid* realizations of $\mathcal{M}_z$. We assume that the population $\mathcal{F}$ is sampled according to a sample design $p(\mathbf{S}=\mathbf{s})$ where $\mathbf{S}$ is the random vector for the sample membership indicator defined by $\mathbb{E}(\mathbf{S})=\boldsymbol{\pi}$ and $\mathbb{V}(\mathbf{S})=\boldsymbol{\Delta}$. We are interested in computing the population total of y, Y, using the auxiliary variables $\mathbf{x}$ observed in the sample and the known population totals $(N,\mathbf{X})$.

We can take advantage of the relationship between y_k and $\mathbf{x}_k$ described in $\mathcal{M}_z$ by assuming a working model for y_k conditioned on the observed values $\mathbf{x}_k$, and $y_k \mid \mathbf{x}_k$. Since $\mathcal{M}_z$ is a multivariate normal distribution, the conditional distribution of $y_k \mid \mathbf{x}_k$ is a univariate normal distribution (Casella & Berger, 2002) with the parameters

$$y_k \mid \mathbf{x}_k \sim \mathcal{N}\left(\theta_\beta, \theta_\sigma^2\right), \tag{1.49}$$

where $\theta_\beta = \beta_y + \boldsymbol{\Sigma}_{y\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}(\mathbf{x}-\boldsymbol{\beta}_{\mathbf{x}})$ and $\theta_\sigma^2 = \sigma_{y_1}^2\boldsymbol{\Sigma}_{y\mathbf{x}} - \boldsymbol{\Sigma}_{y\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}y}$. We proceed in the same way as before to derive the PA adjusted fitted means $\hat{\mu}_{pa,k}$ by solving the PL of the distribution of $y_k \mid \mathbf{x}_k$ and the observed data to obtain the PMLE of the model mean $\hat{\theta}_{pmle,\beta} = \hat{\mu}_{pmle,k}$ consisting of the PML estimators listed in Table 1.14.

Plugging the PML estimators and the PA adjustments $\hat{\Gamma}_1 = \frac{N}{\hat{N}_{HT}}$ and $\hat{\mathbf{\Gamma}}_{\mathbf{X}} = \mathbf{D}_{\mathbf{X}}\mathbf{D}_{\hat{\mathbf{X}}}^{-1}$ into the generic expression of the PA estimator in (1.25), and after algebraic simplification, the PA estimator for the total Y is

$$\hat{Y}_{PA} = N\hat{\bar{Y}}_{HJ} + \left(\mathbf{X} - N\hat{\bar{\mathbf{X}}}_{HJ}\right)\hat{\mathbf{\Sigma}}_{\mathbf{e_{xx}}}^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{e}_{\mathbf{x}y}}, \tag{1.50}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{e_{xx}}}$ is the design-based estimate of the variance-covariance matrix $\mathbb{C}(\mathbf{e_x}) \in \mathbb{R}^{P\times P}$ of the auxiliary variable residuals $\mathbf{e}_{k\mathbf{x}} = \mathbf{x}_k - \hat{\bar{\mathbf{X}}}_{HJ}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{e}_{\mathbf{x}y}}$ is the design-based estimate of the covariance vector $\mathbb{C}\left(\mathbf{e_x}, \mathbf{e_{xy}}\right) \in \mathbb{R}^{P\times 1}$ between the residuals $\mathbf{e}_{k\mathbf{x}}$ and $\mathbf{e_y} = \mathbf{y} - \hat{\bar{Y}}_{HJ}$. The PA estimator in 1.50) exists if $\hat{\mathbf{\Sigma}}_{\mathbf{xx}}$ is invertible (e.g., full rank, $\operatorname{rank}\hat{\mathbf{\Sigma}}_{\mathbf{e_xe_x}} = P$). The expression in (1.50) is new and has not been previously reported in the literature as far as we know.

Suppose we use the central multivariate normal distribution to produce another estimator. The central multivariate normal distribution has the same expression as above but with zero vector means, $\left(\beta_y, \boldsymbol{\beta}_{\mathbf{x}}^{\mathrm{T}}\right) = \left(0, \mathbf{0}_P\right)$. We proceed in the same way as before to derive the PA adjusted fitted means $\hat{\mu}_{pa,k}$ by solving the PL to obtain the PML estimator of the model mean $\hat{\mu}_{pmle,k} = \hat{\theta}_{pmle,\beta}$ consisting of the PML estimators listed in Table 1.15.

Table 1.14 PMLE of the components of $\hat{\mu}_{pmle,k}$ in Example 1.14

PML Estimator	Expression	Notes
$\hat{\beta}_{pmle,y}$	$\bar{\hat{Y}}_{HJ}$	$\bar{Y}_{HJ} = \left(\mathbf{d}^{\mathrm{T}}\mathbf{y}\right) / \left(\mathbf{d}^{\mathrm{T}}\mathbf{1}\right)$, $\mathbf{d} = \left[d_k\right] \in \mathbb{R}^{A\times 1}$
$\hat{\boldsymbol{\beta}}_{pmle,x}$	$\bar{\hat{\mathbf{X}}}_{HJ}$	$\bar{\hat{\mathbf{X}}}_{HJ} = \mathbf{d}^{\mathrm{T}}\mathbf{x} / \left(\mathbf{d}^{\mathrm{T}}\mathbf{1}\right)$
$\hat{\mathbf{x}}_{pmle}$	$\hat{\mathbf{X}}_{HT}$	$\hat{\mathbf{X}}_{HJ} = \mathbf{d}^{\mathrm{T}}\mathbf{x}$
$\hat{\boldsymbol{\Sigma}}_{pmle,\mathbf{xx}}$	$\hat{\boldsymbol{\Sigma}}_{\mathbf{e}_{\mathbf{xx}}} = \left(\mathbf{d} \odot \breve{\mathbf{e}}_{\mathbf{x}}\right)^{\mathrm{T}} \breve{\boldsymbol{\Delta}} \left(\mathbf{d} \odot \breve{\mathbf{e}}_{\mathbf{x}}\right) \hat{N}_{HT}^{-1}$	$\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$, $\breve{\mathbf{e}}_{\mathbf{x}} = \left(\mathbf{x} - \bar{\hat{\mathbf{X}}}_{HJ}\right) \odot \mathbf{S}$ and $\hat{N}_{HT} = \mathbf{d}^{\mathrm{T}}\mathbf{1}$
$\hat{\boldsymbol{\Sigma}}_{pmle,y\mathbf{x}}$	$\hat{\boldsymbol{\Sigma}}_{\mathbf{e}_{\mathbf{x}y}} = \left(\mathbf{d} \odot \breve{\mathbf{e}}_{\mathbf{x}}\right)^{\mathrm{T}} \hat{\boldsymbol{\Delta}} \left(\mathbf{d} \odot \breve{\mathbf{e}}_{\mathbf{y}}\right) \hat{N}_{HT}^{-1}$	$\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$, $\breve{\mathbf{e}}_{\mathbf{y}} = \left(\mathbf{y} - \bar{\hat{Y}}_{HJ}\right) \odot \mathbf{S}$ and $\hat{N}_{HT} = \mathbf{d}^{\mathrm{T}}\mathbf{1}$

Table 1.15 PMLE of the components of $\hat{\mu}_{pmle,k}$ of the noncentral working model in Example 1.14

PML Estimator	Expression	Notes
$\hat{\beta}_{pmle,y}$	$\hat{Y}_{HT}$	$\hat{Y}_{HT} = \mathbf{d}^{\mathrm{T}}\mathbf{y}$, $\mathbf{d} = \left[d_k\right] \in \mathbb{R}^{A\times 1}$
$\hat{\boldsymbol{\beta}}_{pmle,x}$	$\hat{\mathbf{X}}_{HT}$	$\hat{\mathbf{X}}_{HT} = \mathbf{d}^{\mathrm{T}}\mathbf{x}$
$\hat{\mathbf{x}}_{pmle}$	$\hat{\mathbf{X}}_{HT}$	$\hat{\mathbf{X}}_{HJ} = \mathbf{d}^{\mathrm{T}}\mathbf{x}$
$\hat{\boldsymbol{\Sigma}}_{pmle,\mathbf{xx}}$	$\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}} = \left(\mathbf{d} \odot \mathbf{x} \odot \mathbf{s}\right)^{\mathrm{T}} \hat{\boldsymbol{\Delta}} \left(\mathbf{d} \odot \mathbf{x} \odot \mathbf{s}\right) \hat{N}_{HT}^{-1}$	$\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$, $\hat{N}_{HT} = \mathbf{d}^{\mathrm{T}}\mathbf{1}$
$\hat{\boldsymbol{\Sigma}}_{pmle,y\mathbf{x}}$	$\hat{\boldsymbol{\Sigma}}_{\mathbf{x}y} = \left(\mathbf{d} \odot \mathbf{x} \odot \mathbf{s}\right)^{\mathrm{T}} \boldsymbol{\Delta} \left(\mathbf{d} \odot \mathbf{y} \odot \mathbf{s}\right) \hat{N}_{HT}^{-1}$	$\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$, $\hat{N}_{HT} = \mathbf{d}^{\mathrm{T}}\mathbf{1}$

Plugging the PMLE estimators and PA adjustments $\hat{\mathbf{\Gamma}}_{\mathbf{X}} = \mathbf{D}_{\mathbf{X}}\mathbf{D}_{\hat{\mathbf{X}}}^{-1}$ into the generic expression of the PA estimator in (1.25), and after algebraic simplification, the PA estimator for the population total Y based on the central multivariate normal distribution is

$$\hat{Y}_{PA} = \hat{Y}_{HT} + \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right)\hat{\mathbf{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{x}y}, \tag{1.51}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{xx}} \in \mathbb{R}^{P\times P}$ is the design-based estimate of the variance-covariance matrix $\mathbb{C}\left(\hat{\mathbf{X}}_{HT}\right) \in \mathbb{R}^{P\times P}$, and $\hat{\mathbf{\Sigma}}_{\mathbf{x}y} \in \mathbb{R}^{P\times 1}$ is the design-based estimate of the variance-covariance vector $\mathbb{C}\left(\hat{\mathbf{X}}_{HT}, \hat{\mathbf{y}}_{HT}\right) \in \mathbb{R}^{P\times 1}$. The estimator (1.51) exists if $\hat{\mathbf{\Sigma}}_{\mathbf{xx}}$ is invertible, e.g. $\operatorname{rank}\hat{\mathbf{\Sigma}}_{\mathbf{xx}} = P$.

The estimator (1.51) is the *Randomization Optimal Estimator* proposed by Montanari (1987, 1998, and 2002) that has been extensively studied in the literature (Fuller & Isaki, 1981; Cassady & Valiant, 1993; Rao, 1994; Tillé, 1999; Chen & Sitter, 1999; and Montanari & Ranalli, 2002).

We refer to the estimator in (1.51) as the central optimal estimator and (1.50) is the noncentral optimal estimator. For survey data where the outcome variable and the auxiliary variables are positive, the model for the noncentral optimal estimator is misspecified since the parameters means $\boldsymbol{\beta}$ are not generally zero. However, as a model-assisted estimator, (1.51) is still design consistent. In contrast, the working model of noncentral optimal estimators is more plausible because the means do not have to be zero in the working model. The differences between the estimators are that

(1.50) uses the HJ estimators of $\bar{Y}$ and $\bar{\mathbf{X}}$ while in (1.51) the HT estimators of Y and $\mathbf{X}$ are used. The variance-covariance matrix in (1.50) is based on the estimated total residuals $\sum_{k\in A} d_k \mathbf{e}_k = \sum_{k\in A} d_k \left(\mathbf{x}_k - \hat{\bar{\mathbf{X}}}_{HJ}\right)$ while (1.51) is based on the estimated totals $\hat{\mathbf{X}}_{HT}$. We hypothesize that gains in efficiency of the optimal estimator are due to the type of model because this model describes the correlation among all auxiliary variables and the outcome variable.

We do not include an evaluation of the non-central optimal estimator, but it is expected to be more efficient than the central optimal estimator when the HJ estimators for the auxiliary variables have a better fit to the data. One difficulty in fitting the central and non-central optimal estimators under the PA approach is the selection of the auxiliary variables of the working model. These models are not fitted using standard functions for generalized linear regression models and require developing specialized routines for computing and maximizing the PL functions for this type of model.

1.7.6 The Horvitz-Thompson Estimator

The HT estimator is referred to as the only true model-free design-based estimator; it is a "no information" estimator in the sense that no population totals are used[10]. The

[10] The HT model described in Chen, et al. (2017) is used to predict non-sampled cases and differs from the "no information" view of the HT estimator.

HT estimator results from any working model (linear or nonlinear) without any PA adjustment (equivalent to a PA adjustment $\hat{\mathbf{\Gamma}}_{\mathbf{X}} = \mathbf{D}_{\mathbf{X}}\mathbf{D}_{\hat{\mathbf{X}}}^{-1} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix). These results can be summarized in the following theorem:

THEOREM 1.6. Let $\hat{Y}_{\mu} = \sum_{k \in A} d_k \hat{\mu}_{PML,k}$ be an estimator consisting of the sum of the expanded values of fitted PMLE of the means of the assumed working model $\mathcal{M}_y$, then $\hat{Y}_{HT} - \hat{Y}_{\mu} = 0$ and $\mathrm{V}\left(\hat{Y}_{HT}\right) - \mathrm{V}\left(\hat{Y}_{\mu}\right) = 0$.

In other words, the estimator based on fitted means of a working model without any auxiliary variable is the same as the HT estimator. There are no gains in efficiency by fitting a model without any population totals.

1.8 Auxiliary Variables and Population Totals

Within the PA framework, we define the auxiliary variables as $\mathbf{x}_k \in \mathbb{R}^{1\times P}$ for $k \in A$ where the population totals $\mathbf{X}$ are known[11]. For the PA estimators in this paper, the additional information from the auxiliary variables consists only of the population totals $\mathbf{X}$. If complete auxiliary information is available (i.e., the values of $\mathbf{x}_k$ are

[11] Other classes of PA such as those that require complete auxiliary information or estimators that incorporate estimated population totals from the sample are not described in this dissertation.

known for every $k \in U$), it is summarized to produce population totals. The population totals are considered fixed.

We consider two types of auxiliary variables. The first group includes the sample design variables, that is, those variables created at the design stage or used to select the sample. We list seven of these types of auxiliaries:

1. Unit auxiliary variable. The simplest auxiliary variable is a vector with a value of one for all members of the population; the population total is N. The unit auxiliary variable allows an intercept term in the regression model of the parameters of the working model; this allows ML and PML models such that the sum of the residuals and weighted residuals are asymptotically zero for valid PA models.

2. First order probabilities of inclusion π_k for $k \in A$ with a population total $n = \sum_{k \in U} \pi_k$ that corresponds to the expected sample size. For sample designs where $\pi_k \propto x_k$, both variables are equivalent since one is the scaled version of the other.

3. Sample design weights $d_k = \pi_k^{-1}$ for $k \in A$ with a population total $D_U = \sum_{k \in U} d_k$.

   The sample design weights can be scaled for numerical stability when maximizing the PML function. Using the weight as an auxiliary variable requires complete information on the weights to compute the population total D_U.

4. Certainty indicator. The indicator c_k that identifies if a sample unit is selected with certainty, $c_k = 1$, or $c_k = 0$, otherwise. The population total is the number of cases sampled with certainty.

5. Stratum membership indicator defined as the vector $\mathbf{h}_k = \left(h_{k1},...,h_{kh'},...,h_{kH}\right) \in \{0,1\}^{1\times H}$ with $h' \in \{1,...,H\}$ for $k \in A$ where H is the number of strata, and $h_{kh'} = 1$ if the element k is in stratum h', and $\mathrm{h}_{kh'} = 0$ otherwise. The population total is $\mathbf{H} = \left(H_1,...,H_h,...,H_H\right) \in \mathbb{R}^{H\times 1}$ where $\mathbf{H} = \sum_{k\in U} \mathbf{h}_k$.

6. Qualitative or categorical auxiliary variables are defined by a vector of group membership indicators $\mathbf{g}_k = \left(g_{k1},...,g_{kg'},...,g_{kG}\right) \in \{0,1\}^{1\times G}$ with $g' \in \{1,...,G\}$ for $k \in A$ where G is the number of groups or categories, and $g_{kg'} = 1$ if the element k is in group g' and $g_{kg'} = 0$, otherwise. The population total is $\mathbf{G} = \left(G_1,...,G_{g'},...,G_G\right) \in \mathbb{R}^{G\times 1}$ where $\mathbf{G} = \sum_{k\in U} \mathbf{g}_k$. Examples of categorical variables are gender, age groups, or geographic areas that are very common in population surveys (Brick, 2013).

7. Quantitative or continuous auxiliary variables. This type of auxiliary variable is commonly found in establishment surveys but is rare in population surveys. Some examples of quantitative auxiliary variables are the total number of

patients seen during a period, the number of doctor visits at the end of a period, taxable income, or total revenue.

Additional auxiliary variables can be derived from the interaction of the quantitative, qualitative, and sampling variables. For example, the unit sample indicator and continuous variables can produce the regression or multiple regression estimators, or the interaction between sampling stratum indicators and a continuous variable yield to the separate ratio estimator.

EXAMPLE 1.15. Assume two vectors of auxiliary variables $\mathbf{g}_k$ and $\mathbf{g}'_k$ with the membership indicator for the levels of two categorical variables G_1, G_2, and a PA fully saturated linear model for the outcome variable with a normal distribution $y_k \overset{iid}{\sim} \mathcal{N}\left(\beta_g + \beta_{g'} + \beta_{g \cdot g'}, \sigma^2_{g \cdot g'}\right)$ for $g \in \{1,...,G\}$ and $g' \in \{1,...,G'\}$. This model corresponds to the cross-tabulation of $\mathbf{g}$ and $\mathbf{g}'$ with , β_g (rows) and $\beta_{g'}$ (columns) as main effects, and the interaction term $\beta_{g \cdot g'} = \beta_g * \beta_{g'}$. We assume that the population totals $\mathbf{G} * \mathbf{G}' = \left(N_{11},...,N_{GG'}\right)$ are available. Table 1.16 lists four and PA estimators with different working models depending on the fit of the data. The first PA estimator is the canonical HJ estimator for the single mean model where there are no differences among the means of the cells $\mathbf{g} \cdot \mathbf{g}'$. The second and third estimators are for the main effect models ($\mathbf{g}$ or $\mathbf{g}'$) where there are no differences in the means among columns (estimator 2) or rows (estimator 3) among columns. The last

estimator is for the fully saturated model where there are differences among the means of the cells $\mathbf{g} \cdot \mathbf{g}'$.

Table 1.16 PA estimators of Example 1.15

Model	PA Estimator	Notes
1. Single mean	$\hat{Y}_{PA} = \frac{\hat{Y}_{HT}}{\hat{N}_{HT}} N$	$\hat{Y}_{HT} = \sum_{g \in G} \sum_{g' \in G'} \sum_{k \in a_{gg'k}} d_{gg'k} y_{gg'k}$, $\hat{N}_{HT} = \sum_{g \in G} \sum_{g' \in G'} \sum_{k \in a_{gg'k}} d_{gg'k}$, $N = \sum_{g \in G} \sum_{g' \in G'} N_{gg'}$
2. Row main effects **g**	$\hat{Y}_{PA} = \sum_{g \in G} \frac{\hat{Y}_{HT,g}}{\hat{N}_{HT,g}} N_g$	$\hat{Y}_{HT,g} = \sum_{g' \in G'} \sum_{k \in a_{gg'k}} d_{gg'k} y_{gg'k}$, $\hat{N}_{HT,g} = \sum_{g' \in G'} \sum_{k \in a_{gg'k}} d_{gg'k}$, $N_g = \sum_{g' \in G'} N_{gg'}$
3. Column main effects **g′**	$\hat{Y}_{PA} = \sum_{g' \in G'} \frac{\hat{Y}_{HT,g'}}{\hat{N}_{HT,g'}} N_{g'}$	$\hat{Y}_{HT,g'} = \sum_{g \in G} \sum_{k \in a_{gg'k}} d_{gg'k} y_{gg'k}$, $\hat{N}_{HT,g'} = \sum_{g \in G} \sum_{k \in a_{gg'k}} d_{gg'k}$, $N_{g'} = \sum_{g \in G} N_{gg'}$
4. Fully saturated **gg′**	$\hat{Y}_{PA} = \sum_{g \in G} \sum_{g \in G'} \frac{\hat{Y}_{HT,gg'}}{\hat{N}_{HT,gg'}} N_{HT,gg'}$	$\hat{Y}_{HT,gg'} = \sum_{k \in a_{gg'k}} d_{gg'k} y_{gg'k}$, $\hat{N}_{HT,gg'} = \sum_{k \in a_{gg'k}} d_{gg'k}$

REMARK 1.16 Little (2008) discusses the model-based estimation for the setting where $\mathbf{g}$ are the strata and $\mathbf{g}'$ are the poststrata, and the saturated model is replaced by an additive model with main effects for strata and poststrata when the stratum/poststratum cells have few observations. The PA estimator adopts a prediction perspective that corrects the usual poststratified estimator based only on $\mathbf{g}'$ so it can produce estimators that match both stratum and post-stratum margins while allowing modifications of the fully saturated estimator in small samples by modifying the distribution of the cell means. The effect of replacing the saturated model by the simpler main effects model shrinks the estimates of the stratum/poststratum cell sample means of the saturated model towards the means of the additive model. The shrinkage of the sample means occurs during working model development in the PA where simpler working models with a lower loss function replace the complex model in the algorithm. Little (2008) describes this shrinkage of post-stratum means as a desirable property of an estimator from the modeling perspective. In the extreme case, when the optimal model has only one stratum, the initial model sample means shrink towards the overall mean, which corresponds to the canonical form of the HT estimator.

REMARK 1.17 If there is one categorical auxiliary variable for the poststratification cells, the algorithmic PA estimator can be used for collapsing poststrata without modifications to the algorithm.

REMARK 1.18 Särndal & Lundström (2005) describe the unit auxiliary variable as the simplest auxiliary vector that does not recognize individual differences among the elements of the population. If we assume a normal linear model for y, $y_k \overset{iid}{\sim} \mathcal{N}\left(\beta, \sigma_0^2\right)$ with the auxiliary variable 1 and a population total N, the PMLE of the regression coefficient is $\hat{\beta}_{pmle} = \frac{\hat{Y}_{HT}}{\hat{N}_{HT}}$, the PA adjusted regression coefficient is $\hat{\beta}_{pa} = N\frac{\hat{Y}_{HT}}{\hat{N}_{HT}^2}$, and the PA estimator is $\hat{Y}_{PA} = N\frac{\hat{Y}_{HT}}{\hat{N}_{HT}}$ which matches the canonical form of the HJ estimator (see Definition 1.2).

EXAMPLE 1.16. In this example, we examine the effect on the efficiency when the variances are modeled in the PA estimator. The documentation of the command `svyglm` in the package `survey` (Lumley, 2012) shows an example for computing three estimates and their variances for the total number of students tested (variable `api.stu`) using a continuous variable with the school's student enrollment (variable `enroll`) from the data file `api` for the Academic Performance Index (API) for all California schools. In this design, the frame consists of 6,157 California schools stratified by school type with 4,397 elementary schools, 1,009 middle

schools, and 751 high schools.[12] A total sample of 200 schools is disproportionally allocated to the three strata and three independently simple random samples of 100 elementary schools, 100 middle schools, and 50 high schools are drawn from each stratum.

Lumley (2012) produces three estimators: the GREG estimates, and two ratio estimators with a variance as a function of the mean (μ and μ^3) listed in Table 1.17. Since this ratio estimator with a variance as a function of μ^3 has a smaller standard error for the observed sample, Lumley states that a higher efficiency is achieved by better modeling the variance. The last row of the table shows the algorithmic PA estimator for the same sample, with an assumed working models $y_k \overset{iid}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_k, \sigma_0^2\right)$ where only the location parameter of the distribution is modeled.

The relative efficiency of the estimators for repeated sampling is shown in Table 1.18 for $B =$ 100,000 draws (See Section A.4 in Appendix A for the definitions of the empirical measures of precision in Monte Carlo studies). The results show that although all estimators are more efficient than the HT estimator (12 times more efficient), the gains in efficiency are relatively small when the variance is explicitly

[12] The data file `apipop` in Lumley (2012) contains 6,194 schools. There are 35 schools with missing values of the variable `enroll`. The variable `enroll` is used to compute the total X for the ratio estimators. Those schools with missing values were removed from the file before the simulation and when computing the estimates in Table 1.12.

modeled. In other words, the reduction in standard errors when modeling the variance as μ^3 in Table 1.17 is not typical under repeated sampling. Note that the algorithmic PA estimator does not achieve the largest RE, but the difference with respect to the largest value is less than one percentage point.

Table 1.17 Population totals, estimates, and standard errors for the total number of students tested for three models from Lumley (2012) and two algorithmic PA estimators

Population variable	Description		Total
Schools	Number of California schools in frame		6,157
Enrolled students	Total enrolled students in CA schools in frame		3,8114,72
API Students	API students tested in CA schools in frame		3,184,662
Estimators of total API students	Working Models	Estimates	Standard error
1. GREG	$\mathcal{N}\left(\beta_0+\beta_1 x_k, \sigma_0^2\right)$	3,186,758	31,341
2. Ratio estimator - μ	$\mathcal{N}\left(\beta_1 x_k, x_k \sigma_0^2\right)$	3,190,038	29,566
3. Ratio - μ^3	$\mathcal{N}\left(\beta_1 x_k, x_k^3 \sigma_0^2\right)$	3,247,986	21,129
4. Algorithmic PA	$\mathcal{N}\left(\beta_0+\beta_1 x_k, \sigma_0^2\right)$	3,196,977	28,636

Table 1.18 Empirical summary results for 100,000 draws for Example 1.11.

Estimator	Relative Bias (RB) (%)	Relative Root Mean Squared Error (RRMSE)	Relative efficiency (RE)
1. HT	0.02	3.406	0.00
2. GREG	-0.01	0.939	12.15
3. Ratio - $\sigma^2 \propto \mu$	0.00	0.919	12.73
4. Ratio - $\sigma^2 \propto \mu^3$	0.00	0.914	12.87
5. Algorithmic PA	-0.04	0.918	12.77

EXAMPLE 1.17. Lumley, Shaw, & Dai (2011) provide an example of a more complex auxiliary variable derived from the frame that can be used in the PA working models. Their variable is based on the empirical influence function of a multiple linear regression model. The influence function of a parameter describes the effect on the estimator when changing one point of the data. After identifying a variable with a strong linear relationship with the outcome z, a linear model is fit using P explanatory variables available in the frame as $\hat{z}_k = \hat{\boldsymbol{\beta}}\mathbf{x}_k$, where $\mathbf{x}_k = (x_{k1},\ldots,x_{kP}) \in \mathbb{R}^{1\times P}$ are the auxiliary variables and $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1,\ldots,\hat{\beta}_P\right)^{\mathrm{T}} \in \mathbb{R}^{P\times 1}$ are the fitted regression coefficients. Let $\boldsymbol{\mathcal{I}}_k = \left(\mathcal{I}_{k1},...,\mathcal{I}_{kP}\right)$ be the vector with the values of the empirical influence function of each regression coefficient of a fitted regression for $k \in U$. The vector of the auxiliary variables is $\mathbf{x}_k = \left(\mathbf{j}_P + \boldsymbol{\mathcal{I}}_k\right)$ where $\mathbf{j}_P \in \mathbb{R}^P$ is the one vector $\mathbf{j}_P = \left(1,...,1_P\right)$, and the population total is $\mathbf{X} \in \mathbb{R}^P$ where

$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k = (N, ..., N)$. Since the population totals of the values of the empirical influence function are zeros, a value of one is added to each variable $\mathcal{I}_k$ to ensure that the PA adjustment $\hat{\Gamma}_{\mathcal{I}_p} = \frac{N}{\hat{N}_{HT} + \hat{\mathcal{I}}_{HT,p}}$ is not undefined. Note that even if this auxiliary variable is derived from a model, the variance-covariance of the population totals is zero, e.g., $\mathbb{C}(\mathbf{X}) = \mathbf{0}$.

Chapter 2 The Applications of Algorithmic PA Estimators

In this chapter we describe three applications of PA algorithmic estimators. In the first, we show how the PA framework is used to select the auxiliary variables for the working model of the estimator. In the second, we evaluate linear and nonlinear algorithmic estimators derived using the PA framework. In the last example, we derive and evaluate two algorithmic estimators in samples from Poisson sample designs. Both estimators share the same auxiliary variables, but one has a more complex working model with different regressions for location and scale parameters.

2.1 Variable Selection for Calibration Estimators

The most important application of the PA framework is the selection of variables for calibration estimators in the presence of full response. As noted by Kott (2016), Kott & Liao (2017), and Valliant, Dever, & Kreuter (2013), there is limited work on the methodology for developing working models for model-assisted estimators within the design-based context. Ruppert (2007) and Opsomer, Breidt, Moisen, & Kauermann (2007) share similar views and highlight the need for methods for variable selection in model-assisted estimators. For example, these methods are needed to identify situations where the model-assisted estimator is less efficient than simple estimators such as the Horvitz-Thompson (HT) estimator.

Chambers & Skinner (1999) proposed the creation of weights calibrated to as many auxiliary variables as possible, but this approach is mainly intended for systems of

weights for the analysis of multipurpose surveys (Haziza & Beaumont, 2017). Including auxiliary variables that are not related to the outcome may increase the variability of the weights.

Nascimiento Silva & Skinner (1997) proposed a stepwise method for variable selection based on the mean squared error (MSE) of the linear regression estimator for simple random sampling (SRS) designs. They empirically showed that calibrating to a reduced set of auxiliary variables correlated to the outcome achieves larger gains in efficiency compared to calibrating to a larger set including unrelated variables. However, their approach has severe limitations because their variable selection procedure and expression for the estimate of variance do not generalize beyond SRS designs.

More recently, McConville, Breidt, Lee, & Moisen (2017), denoted as MBLM henceforth, proposed a model-assisted estimator for population totals based on the Least Absolute Shrinkage and Selection Operator (LASSO) developed by Tibshirani (1996). The LASSO is a regression analysis method that performs both variable selection and regularization that improves the prediction accuracy and interpretability of the model. In the LASSO variable selection process, the explanatory variables associated with regression coefficients with small or zero values are eliminated from the initial model. From the PA framework viewpoint, although the superpopulation model $\mathcal{M}_y$ for y is $y_k \overset{iid}{\sim} \mathcal{N}\left(\mu_k, \sigma_0^2\right)$, the procedure fits $\hat{\mu}_{LASSO0,k} = \mathbf{x}_k \hat{\boldsymbol{\beta}}_{LASSO0}$, where $\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kP}\right) \in \mathbb{R}^{1\times P}$ for $k \in U$ is the vector of the auxiliary variables

associated with the LASSO regression coefficients $\hat{\boldsymbol{\beta}}_{LASSO} = \left(\beta_1,\ldots,\beta_{P'}\right)^{\mathrm{T}} \in \mathbb{R}^{P'\times 1}$ computed as

$$\hat{\boldsymbol{\beta}}_{LASSO} = \underset{\boldsymbol{\beta}\in\mathbf{B}}{\arg\min}\left\{\sum_{k\in A} d_k\left(y_k - \boldsymbol{\beta}\mathbf{x}_k\right)^2\right\} \text{ subject to } \left\|\boldsymbol{\beta}\right\|_1 < t\,, \tag{2.1}$$

where t is a prespecified parameter that determines the amount of regularization, d_k are the sampling weights, and $\left\|\boldsymbol{\beta}\right\|_1$ is the $L-1$ norm of the parameter vector $\boldsymbol{\beta}\in\mathbb{R}^{P\times 1}$ such as $P'\subseteq P$. The population total for $\hat{\mu}_{LASS0,k}$ is $\widetilde{M} = \sum_{k\in U}\hat{\mu}_{LASS0,k} = \mathbf{X}\hat{\boldsymbol{\beta}}_{LASS0}$. The expression of the MBLM estimator of the total Y computed using the auxiliary variables $\hat{\mu}_{LASS0,k}$, and the population total $\widetilde{M}$ is

$$\begin{aligned}\hat{Y}_{LASSO} &= \sum_{k\in A} d_k y_k + \widetilde{M} - \sum_{k\in A} d_k\hat{\mu}_{LASS0,k} \\ &= \hat{Y}_{HT} + \left(\widetilde{M} - \hat{M}_{HT}\right)\hat{\boldsymbol{\beta}}_{LASS0}\end{aligned}\,. \tag{2.2}$$

Although the method for producing the LASSO estimator can be used to select variables of the working model, the method does not produce a calibration estimator in the sense that the calibrated weights meet the calibration equations (Deville & Särndal, 1992; Deville, Särndal, & Sautory 1993). MBLM derives a calibration estimator using a secondary working model $y_k \overset{iid}{\sim} \mathcal{N}\left(\hat{\mu}_{LASS0,k}\alpha_0 + \mathbf{x}_k\boldsymbol{\alpha}_{\mathbf{x}}, \sigma_0^2\right)$ with auxiliary variables $\mathbf{x}_k^* = \left(\hat{\mu}_{LASS0,k}, \mathbf{x}_k\right)$ and population totals $\mathbf{X}^* = \left(\widetilde{M}, \mathbf{X}\right)$. The calibration LASSO estimator is

$$\hat{Y}_{cal\_LASSO} = \hat{Y}_{HT} + \left(\widetilde{M} - \hat{M}\right)\hat{\alpha}_0 + \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right)\hat{\boldsymbol{\alpha}}_{\mathbf{x}}.$$

where $\hat{\boldsymbol{\alpha}} = \left(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}_{\mathbf{x}}^{\mathrm{T}}\right)^{\mathrm{T}}$ computed as $\hat{\boldsymbol{\alpha}} = \hat{\mathbf{T}}_{\mathbf{x}^*\mathbf{x}^*}^{-1}\hat{\mathbf{T}}_{\mathbf{x}^*\mathbf{y}}$, where $\hat{\mathbf{T}}_{\mathbf{x}^*\mathbf{x}^*} = \sum_{k\in A} d_k \left(\mathbf{x}_k^*\right)^{\mathrm{T}} \mathbf{x}_k^*$ and $\hat{\mathbf{T}}_{\mathbf{x}^*\mathbf{y}} = \sum_{k\in A} d_k \left(\mathbf{x}_k^*\right)^{\mathrm{T}} y_k$. We propose a modification to the LASSO procedure that calibrates to the auxiliary variables of the model identified by the LASSO procedure instead of calibrating to the total $\widetilde{M} = \sum_{k\in U} \hat{\mu}_{LASS0,k}$. The modified LASSO estimator is a traditional calibration estimator with the relevant auxiliary variables that explain the outcome variable similar to the PA estimator. This modification is an alternative to the PA algorithm but using (2.2) as the loss function. The evaluation of the MBLM estimator and the modified LASSO estimator are not included here; but our initial evaluation pf this loss function suggests that there are potential issues such as the assumption that the model is known, sparse, and well specified.

Chen, Valliant, & Elliott (2018), denoted as CVE henceforth, propose a method for calibrating nonprobability samples to estimated population totals similar to the MBLM estimator, but they use two separate samples and the adaptive LASSO (Zou, 2006). The CVE method does not produce a traditional calibration estimator, but instead gives a GREG estimator with one derived auxiliary variable. The superpopulation model is the same as the MBLM model described above. The derived variable is $\hat{\mu}_{lasso_1,k}$, the estimated mean of the LASSO model fitted to a probability sample A_1 called the analytical sample. The estimated population total of the derived variable $\hat{\mu}_{lasso_1,k}$ is derived as the HT estimator of the predicted means $\hat{\mu}_{lasso_2,k}$ of the LASSO model from the analytical sample but applied to the second

sample A_2 called the benchmark sample. The model identification and variable selection method of the CVE estimator do not apply to estimation from probability samples in the presence of full response that we are considering here.

REMARK 2.1. Fabrizi & Lahiri (2013) proposed a design-based approximation to the Bayes Information Criterion (BIC) in finite population sampling. Although they mentioned the importance of variable selection, they evaluated their design-based BIC using hypothesis of one single parameter of a model because their focus was estimating the parameter of the model rather than the auxiliary variables for the calibration estimator as discussed here. They planned to extend their findings to a general variable selection method but did not give a method that evaluated models based on the design-based BIC.

REMARK 2.2. Pfeffermann & Sverchkov (1999) proposed a likelihood-based method for estimating parameters of models using survey data selected using an informative sampling method. This approach is called sample likelihood (Chambers, Steel, Wang, & Welsh, 2012), and estimates the sample likelihood of parameters of the conditional distribution of the observed data given the auxiliary variables. Their method shares some similarities with PA modelling methods for the sample membership indicators with some important differences: the use of the pseudo-maximum likelihood estimation (PMLE) instead of maximum likelihood estimation (MLE), and the implementation of separate steps for modeling the sampling membership as an outcome variable (Steps 1, 2, 5, and 6 of Algorithm 1.1).

Pfeffermann & Sverchkov (1999) mention that sample likelihood permits the use of standard inference procedures such as MLE or related residual analysis that are building blocks for variable selection methods. However, no method for variable selection or model building based on the sample likelihood has been proposed in the literature.

Sverchkov (2010) extends the sample likelihood approach to estimation in the presence of nonresponse when the probability of responding is related to the outcome variable (e.g., missing data not missing at random or NMAR). As in previous methodology, Sverchkov (2010) notes that the parameters of the models can be estimated by MLE and evaluated using any classical information criteria such as the Akaike AIC or the Schwarz BIC; however, no procedure based on this approach has been reported in the literature. Furthermore, this approach does not address the situation examined in this dissertation, that is, estimation with full response.

REMARK 2.3. It important to note that there a large number of methods for variables selection described in the standard statistical literature. Many new methods based on statistical learning approaches have been developed in recent years. An older review of the standard statistics methods from the frequentist point of view is found in Rao & Wu (2001). Bayesian selection methods are reviewed in Berger & Pericchi (2001); Efron & Gou (2001) attempt to reconcile the frequentist and Bayesian theories with limited success beyond the single parameter setting for the normal distribution.

More recent methods for variable selection, referred as to feature selection within the Machine Learning context, are reviewed by Hastie, Tibshirani, & Friedman (2009) and Somol, Novovicova, & Pudil (2010). One difference between the standard methods and the approach to model selection in Machine Learning is the complete characterization of the algorithms generally not discussed addressed in the standard methods. For example, variable selection methods are classified as wrapper methods (fit a model to a portion of the sample and evaluate using the remaining sample), filter methods (use a measure of error to score subsets of models), or embedded methods (perform feature selection as part of the model building process), They also have specific approaches to the identification and evaluation of models among the full set of possible number (in contrast with few hypothesis tests used in most classical methods). The reason is that this process is time consuming and costly if all models are fitted. The classical and modern methods have their merits, and some of these features are incorporated into the PA variable selection algorithm (e.g., greedy fromward selection with a loss function). However, they all assume that the observed data are independent and identically distributed random variables (*iid*). Furthermore, some methods attempt to minimize the mean squared error (MSE) instead of the bias that is the more common goal in survey estimation. Therefore, most of these methods cannot be imported to the survey sampling context without a theoretical justification or modifications to the procedure to reflect the sample design. As noted in Kott (2016), Kott & Liao (2017), and Valliant, Dever, & Kreuter (2013), there is limited work on the methodology for developing working models for model-assisted

estimators within the design-based context despite the large number of variable selection methods in standard statistics.

REMARK 2.4. One important difference between the standard statistical methods and those based on Machine Learning is the reliance on statistical tests in the former versus the test-free optimality criteria of the latter methods. This difference is key to the role of the estimated parameters of the fitted working model within the PA framework and in survey sampling estimation in general. In the PA approach, the values of the estimated model parameters are not important since no inference is made. This is sensible because the population characteristic such as totals or means should be robust to the values of the parameters of assumed models that are unknown or inestimable. In the PA approach, there is no hypothesis testing or any other statistical measure for each estimated model parameter. Only the fit of the model drives the inclusion of the variables in the model. The model fit affects the residuals of the estimates, which in turn have an impact on the variance. Although the model is important, the goal is not identifying the true model. Instead, the model is just a tool for producing efficient estimators.

2.2 Variable Selection in Algorithmic PA Estimators

In the first part of this example, we evaluate the algorithm for variable selection for the working model of algorithmic PA estimators based on a single realization of the sample. Since the variables in the working model determine the functional form of the

estimator, this example also evaluates the functional form of the algorithmic PA estimator.

The simulation is motivated by the example in Section 7.9.1 of Särndal, Swensson, & Wretman (1992), denoted as SSW henceforth, where the efficiency of multiple regression estimators is compared to simple estimators. The sampling frame is the MU281 population with 1985 administrative data for 281 Sweden municipalities (Tillé & Matei, 2016).[13] The study variable y is RMT85$\times 10^{-4}$, where RMT85 is the municipal tax receipts received in 1985. Two auxiliary variables on the frame are $x_1 =$ CS82, the number of Conservative Party seats in the municipal council in 1982, and $x_2 =$ SS82, the number of Social Democrat Party seats in the municipal council in 1982. SSW fit different regression models on y from the frame and determine that the multiple regression estimator $\hat{Y}_{\text{SSW},\, x1x2}$ with the model $(1, x_1, x_2)$ has the best fit for the population. Through repeated sampling, they verify that $\hat{Y}_{\text{SSW},\, x1x2}$ is the most efficient among other alternative estimators such as the HT, two ratio estimators with auxiliary variables x_1 and x_2, respectively, and two regression estimators with models $\mathcal{M}_{1,y} = (1, x_1)$ and $\mathcal{M}_{2,y} = (1, x_2)$, respectively. This example has pedagogical value but requires knowing the outcome variable for every unit in the frame.

[13] As in the Särndal, Swensson, & Wretman (1992) simulation, the three municipalities with the largest values of municipal tax receipts received in 1985 in the `MU284` population are removed for the sampling frame.

In the first scenario, we recreate this study evaluating the same group of estimators in the SSW simulation in addition to the algorithmic PA estimator, $\hat{Y}_{\text{PA},\, x1x2}$ with the collection of working models $\boldsymbol{\mathcal{M}}_y$ with $y_k \overset{iid}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma^2\right)$ spanned by the auxiliary variables $\mathbf{x} = (1, x_1, x_2)$, assuming that only the population totals $\mathbf{X} = (N, X_1, X_2)$ are known. The collection of working models $\boldsymbol{\mathcal{M}}_y$ of $\hat{Y}_{\text{PA},\, x1x2}$ can reproduce the models of the other estimators evaluated in the original study.

In each simulation run, a SRS sample of 100 municipalities is drawn, and estimates, their estimated variances, and confidence intervals are computed. These statistics are used to compute the empirical relative bias (RB, in percentage), relative root mean squared error (RRMSE), and relative efficiency (RE in percentage) of the estimator compared to the HT (see the definitions of these empirical summary measures in Section A.4 in Appendix A).

The middle panel of Table 2.1 shows the RB, RMSE, RE, the empirical coverage rate for 95% nominal confidence interval coverage (ECR), and the empirical length of ECR (LECR) of the estimators for $B = 100{,}000$ runs for the first scenario. The table also includes Kish's weighting design effect $deff_{kish} = 1 + cv(w)^2$ where $w = \{w_k\}_{k \in A}$ and w_k are the weights assuming that the sampling weights $d_k = \pi_k^{-1}$ are calibrated to the population totals of the model of the algorithmic PA estimator. Table 2.1 shows the results for the estimators $\hat{Y}_{\text{SSW},\, x1x2}$, $\hat{Y}_{\text{PA},\, x1x2}$ and HT; the HT estimator is used as a reference. The empirical bias of the estimators $\hat{Y}_{\text{SSW},\, x1x2}$ and

$\hat{Y}_{\mathrm{PA},\,x1x2}$ are both very small, less than 0.3 percentage points as expected. Both estimators are 2.7 times more efficient than $\hat{Y}_{HT}$. The table shows that the algorithmic PA estimator $\hat{Y}_{\mathrm{PA},\,x1x2}$ is as efficient as the estimator $\hat{Y}_{\mathrm{SSW},\,x1x2}$ identified by SSW, even though $\hat{Y}_{\mathrm{PA},\,x1x2}$ is based on the observed sample in each simulation run. Both estimators $\hat{Y}_{\mathrm{SSW},\,x1x2}$ and $\hat{Y}_{\mathrm{PA},\,x1x2}$ have the same performance because, in each run, the PA algorithm chooses the same model of $\hat{Y}_{\mathrm{SSW},\,x1x2}$, so $\hat{Y}_{\mathrm{PA},\,x1x2} = \hat{Y}_{\mathrm{SSW},\,x1x2}$. In general, the PA algorithm does not necessarily select the same model in all samples, although it does so here.

In the second scenario, we assume there is complete auxiliary information so we can compute a new auxiliary variable $x_3 = x_1 \cdot x_2$ and its population total $X_3 = \sum_{k \in U} x_{k3}$ for the interaction between x_1 and x_2. We compare the algorithmic PA estimator $\hat{Y}_{\mathrm{PA},\,x1x2x3}$ with the collection of working models $\mathcal{M}_y$ spanned by $\mathbf{x} = (1, x_1, x_2, x_3)$ to the multiple regression estimator $\hat{Y}_{\mathrm{SSW},\,x1x2x3}$ with a fixed linear model with the same auxiliary variables $\mathbf{x}$. Note that if the population total X_3 is known, the PA estimator does not require complete auxiliary information data.

The lower pane of Table 2.1 shows the results of the simulation of the second scenario. As in the previous scenario, the estimators have small empirical biases and the estimators $\hat{Y}_{\mathrm{PA},\,x1x2x3}$ and $\hat{Y}_{\mathrm{SSW},\,x1x2x3}$ are 3.5 times more efficient than $\hat{Y}_{HT}$. Using the derived variable x_3 increases the efficiency of the estimators by 20

percentage points over those estimators with a model with only $\mathbf{x} = (1, x_1, x_2)$. A surprising result is that the PA algorithmic estimator $\hat{Y}_{\text{PA},\, x1x2x3}$ is slightly more efficient than $\hat{Y}_{\text{SSW},\, x1x2x3}$ which has no model uncertainty, and the empirical $deff_w$ is smaller for $\hat{Y}_{\text{PA},\, x1x2x3}$ than for $\hat{Y}_{\text{SSW},\, x1x2x3}$. We expected the efficiency of the estimators with a fixed model to be the lower bound of the estimators with uncertainty in their working model.

The ECRs of the estimates in both scenarios are somewhat less than the nominal 95% rate. The more complex estimators (those with four terms in the model) have lower ECRs than the ECR of those with three auxiliary variables). The HT estimator with no auxiliary variables is closer to the nominal coverage. The losses in coverage appear to be due to the complexity of the functional form of the working model, the number of auxiliary variables, and the sample size, and how well variance estimate approximates the variance of the estimate. This effect will be the topic of future research.

While the PA algorithm in Scenario 1 selects only the model $(1, x_1, x_2)$ in all 100,000 runs, in Scenario 2, the algorithm selects only four different models of the 16 possible working models spanned by $(1, x_1, x_2, x_3)$ for $\hat{Y}_{\text{PA},\, x1x2x3}$. Table 2.1 lists the distribution and details of the selected models for the PA estimator $\hat{Y}_{\text{PA},\, x1x2x3}$ in 100,000 simulations runs.

Table 2.1 Results of a simulation for Scenarios 1 and 2 for the example in Section 2.1

Scenario	Estimator	RB (%)	RRMSE $\times 10^2$	RE (%)	ECR	*LECR*	$deff_w$
All	$\hat{Y}_{HT}$	0.036	8.508	0.0	0.938	1.758	1.000
1	$\hat{Y}_{\text{SSW, } x1x2}$	-0.122	4.412	271.9	0.930	0.885	1.014
1	$\hat{Y}_{\text{PA, } x1x2}$	-0.122	4.412	271.9	0.930	0.885	1.014
2	$\hat{Y}_{\text{SSW, } x1x2x3}$	-0.207	4.000	352.5	0.922	0.787	1.022
2	$\hat{Y}_{\text{PA, } x1x2x3}$	-0.169	3.993	354.0	0.922	0.783	1.014

Table 2.2 shows the empirical distribution of the selected working models in $\hat{Y}_{\text{PA, } x1x2x3}$. The probabilities of selecting the models $(1, x_2, x_3)$ and $(1, x_1, x_3)$ are 0.43 and 0.32, respectively. One of these two models is selected about 75 % of the time in repeated sampling. All selected models include the variable x_3, suggesting that this derived variable is more important than x_1 or x_2. In this case, there is no single best model selected for most of the samples.

This example shows that the algorithmic PA estimator is flexible and capable of producing an efficient estimator based on the observed sample. It also shows that the selected auxiliary variables of the final model may vary from sample to sample (Scenario 2) or may be the same for all samples (Scenario 1). The algorithmic PA estimator may be as or more efficient than the estimator with the best model identified when the model for y can be obtained analyzing the full population.

Table 2.2 Empirical distribution of the working models selected by the algorithmic PA estimator $\hat{Y}_{\text{PA, }x1x2x3}$ for 100,000 simulation runs

Estimator	Models $\widehat{\mathcal{M}}_y$				Percentage (%)
	1	x_1	x_2	x_3	
$\hat{Y}_{\text{PA, }x1x2x3}$	✓	✓	✓	✓	6.26
	✓	✓	×	✓	32.40
	✓	×	✓	✓	42.74
	✓	×	×	✓	18.60
Total					100.00

✓: Auxiliary variable selected in the model
×: Auxiliary variable not selected in the model

2.3 Performance of Linear and Nonlinear Algorithmic PA Estimators

In this example, we use simulation to examine the statistical properties of the linear and nonlinear algorithmic PA estimators along with alternative estimators across different types of outcomes, sample designs, and levels of working model misspecification for a range of sample sizes and populations. We evaluate seven estimators for simulation scenarios created by combinations of the factors listed in Table 2.3 for a sequence of 10 populations $\{\mathcal{F}_N\}_{N=1}^{10}$ with increasing sizes $\{N_N\}_{N=1}^{10}$; each sampled at the same rate $f = f_N = \frac{n_N}{N_N}$. Only a subset of these scenarios are presented here, and the full set is presented in Appendix A Section A.1 on page 279.

Table 2.3 Factors in the simulation study for linear and nonlinear PA estimators

Factors	Description
Types of Outcome (3)	1. Bernoulli: Binary data 2. Poisson: Count data 3. Gamma: Continuous positive data with a constant coefficient of variation
Sample designs (3)	1. SRS: Simple random sample without replacement 2. PPS: Probability proportional to size without replacement 3. PO: Poisson sampling
Model strength (3)	1. High 2. Medium 3. Low
Population size (10)	The sequence of populations with increasing size where each population is sampled at the same sampling rate

The available auxiliary variables are $\mathbf{x}_k = (1, x_k)$ with their respective population totals $\mathbf{X} = (N, X)$. These simulations do not evaluate the variable selection of the algorithmic PA because there are only two auxiliary variables.

In this simulation, we examine the numerical performance of algorithmic PA estimators in a setting used to study the estimator's asymptotic properties; that is, through a sequence of increasing population and samples (Isaki & Fuller, 1982; Fuller, 2009). Asymptotic theory does not describe an estimator's performance in small samples, the minimum sample size needed for an estimator to approach its limit, or the performance relative to other estimators (Small, 2010). Since in practice,

the sample size is small in some situations, the numerical results obtained through simulation in this example supplement the asymptotic properties of the PA estimators.

Table 2.4 shows the expressions of seven estimators of the total $Y = \sum_{k \in U} y_k$ evaluated in this simulation study. The first three are the commonly used estimators HT, HJ, and GREG. We include three algorithmic PA estimators: two nonlinear and one linear PA estimator (see Section 1.7.3 on page 85). The first nonlinear PA estimator (NLPA) does not use calibrated weights while in the second (NLCA), the sampling weights are calibrated to the sample size and total population (see Section 1.7.4). The last estimator is the model-calibrated estimator (MC) of Wu & Sitter (2001) described in Remark 1.15. The MC estimator requires auxiliary data for all the elements in the population to be computed. The MC estimator for the Bernoulli population is based on the generalized logistic regression method (GLRE) described in Lehtonen & Veijanen (1998). The new versions of the MC estimator for the Gamma and Poisson populations are derived following the approach in Wu & Sitter (2001), but we include the intercept term.

We use the HT estimator as the reference in the evaluation because it is unbiased for any sample size. In some scenarios, estimators have the same functional form, for example, the HT and the HJ estimators in SRS.

Table 2.4 Seven estimators of the total population Y for the example in Section 2.3 in matrix notation

Estimator	Expression	Notes
1. HT: Horvitz-Thompson	$\hat{Y}_{HT} = \mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{s})$	
2. HJ Hájek	$\hat{Y}_{HJ} = \frac{N}{\hat{N}_{HT}} \mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{s})$	$\hat{N}_{HT} = \mathbf{d}^{\mathrm{T}}\mathbf{s}$
3. GREG: Generalized Regression*	$\hat{Y}_{GREG} = \hat{Y}_{HT} + \left(\mathbf{X} - \hat{\mathbf{X}}_{HT}\right)\hat{\boldsymbol{\beta}}_{ls}$	$\hat{\boldsymbol{\beta}}_{ls} = \hat{\mathbf{T}}_{\mathbf{x},\mathbf{x}}^{-1}\hat{\mathbf{T}}_{\mathbf{x},\mathbf{y}}$ where $\hat{\mathbf{T}}_{\mathbf{x},\mathbf{y}} = (\mathbf{S} \odot \mathbf{d} \odot \mathbf{x})^{\mathrm{T}}\mathbf{y}$ and $\hat{\mathbf{T}}_{\mathbf{x},\mathbf{x}} = (\mathbf{S} \odot \mathbf{d} \odot \mathbf{x})^{\mathrm{T}}\mathbf{x}$
4. NLPA: Algorithmic Nonlinear Parametric	$\hat{Y}_{NLPA} = \mathbf{d}^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_{NLPA} \odot \mathbf{s})$	$\hat{\boldsymbol{\mu}}_{NLPA} = \mathrm{g}^{-1}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pa}\right)$ where $\hat{\boldsymbol{\beta}}_{pa} \in \widehat{\mathcal{M}}_{pa,y} \subset \mathcal{M}_y = (1, x)$ where g^{-1} is logit^{-1} for the Binomial, exp for the Poisson and Gamma populations, and $\hat{\boldsymbol{\beta}}_{pa} = \hat{\boldsymbol{\Gamma}}_{\mathbf{x}}\hat{\boldsymbol{\beta}}_{pmle}$ where $\hat{\boldsymbol{\beta}}_{pmle}$ are the PMLE of a Bernoulli, Poisson, or Gamma distribution model $\mathcal{M}_y$
5. NLCA: Algorithmic Non-linear calibrated PA	$\hat{Y}_{NLCA} = \hat{\mathbf{w}}^{\mathrm{T}}(\mathbf{s} \odot \hat{\boldsymbol{\mu}}_{NLCA})$	Same as NLPA but replacing $\mathbf{d}$ by $\hat{\mathbf{w}}$, the calibrated weights to population totals (n, N).
6. LNPA: Algorithmic Linear Parametric	$\hat{Y}_{LNPA} = \mathbf{d}^{\mathrm{T}}(\mathbf{s} \odot \hat{\boldsymbol{\mu}}_{LNPA})$	$\hat{\boldsymbol{\mu}}_{LNPA} = \mathbf{x}\,\hat{\boldsymbol{\beta}}_{pa}$, $\hat{\boldsymbol{\beta}}_{pa} \in \widehat{\mathcal{M}}_{pa,y} \subset \mathcal{M}_y = (1, x)$, $\hat{\boldsymbol{\beta}}_{pa} = \hat{\boldsymbol{\Gamma}}_{\mathbf{x}}\hat{\boldsymbol{\beta}}_{pmle}$ where $\hat{\boldsymbol{\beta}}_{pmle}$ are the PMLE of a Normal distribution model $\mathcal{M}_y$
7. MC Model Calibrated	$\hat{Y}_{MC} = \hat{Y}_{HT} + \left(\widetilde{\mathbf{M}} - \hat{\mathbf{M}}_{HT}\right)\hat{\boldsymbol{\alpha}}$	With $\mathbf{m} = (1, \hat{\mu}_{mc})$, $\widetilde{\mathbf{M}} = \left(N, \widetilde{M}\right)$, $\hat{\mathbf{M}}_{HT} = \left(\hat{N}_{HT}, \hat{M}_{HT}\right)$, $\widetilde{M} = \mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{\mu}}_{mc}$, $\hat{M} = \mathbf{d}^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_{mc} \odot \mathbf{s})$, $\hat{\boldsymbol{\mu}}_{mc} = \mathrm{g}^{-1}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)$ with g^{-1} is logit^{-1} for the Binomial, exp for the Poisson and Gamma populations, $\hat{\boldsymbol{\beta}}_{pmle}$ are the PMLE of a Bernoulli, Poisson, or Gamma distribution model $\mathcal{M}_y$, $\hat{\boldsymbol{\alpha}} = \hat{\mathbf{T}}_{\mathbf{m},\mathbf{m}}^{-1}\hat{\mathbf{T}}_{\mathbf{m},\mathbf{y}}$ where $\hat{\mathbf{T}}_{\mathbf{m},\mathbf{y}} = (\mathbf{S} \odot \mathbf{d} \odot \mathbf{m})^{\mathrm{T}}\mathbf{y}$ and $\hat{\mathbf{T}}_{\mathbf{m},\mathbf{m}} = (\mathbf{S} \odot \mathbf{d} \odot \mathbf{m})^{\mathrm{T}}\mathbf{m}$

* See Section A.5 in Appendix A on page 304 for the derivation of the linear PA estimator.

The population parameters for the scenarios are listed in Table 2.5. The populations are generated using the linear predictor

$$\eta_k = g(\mu_k) = \beta_0 + \beta_1 x_k + \sigma \varepsilon_k , \tag{2.3}$$

where the parameters β_0, β_1, σ, and the link function $g(\mu_k)$ depend on the scenarios in Table 2.5. The error term ε_k is $\mathcal{N}(0,1)$, and the auxiliary variable x_k has a distribution $\mathcal{B}eta(\alpha,\beta)$ with shape parameters $\alpha = 3$ and $\beta = 6$. For the PPS and PO sample designs, the auxiliary variable x_k is used as the measure of size (MOS) to compute the inclusion probabilities $\pi_k = n\frac{x_k}{\sum_{k\in U} x_k}$, where n is the sample size for the PPS design or the expected sample size for the PO design.

The strength of the model or model misspecification is measured by $\rho_{\eta x}$, the correlation between η_k and x_k, which is a function of σ in the linear predictor η_k. For a fixed value of $\rho_{\eta x}$, $\sigma = \sqrt{\beta_1^2 \operatorname{var}(x)\left(\rho_{\eta x}^{-2} - 1\right)}$. A value of $\rho_{\eta x} =$ 0.9 (high) describes a strong linear relationship between $g(\mu_k)$ and x_k. In this case, we have a well-specified model. The other scenarios are for $\rho_{\eta x} =$ 0.2 (low) and $\rho_{\eta x} =$ 0.6 (medium). Where the relationship is weak or medium, the model is misspecified.

Figure 2.1 shows the scatter plot of the populations of size 10,000 from scenarios in Table 2.5.

Table 2.5 Population parameters and empirical population statistics by simulation scenarios

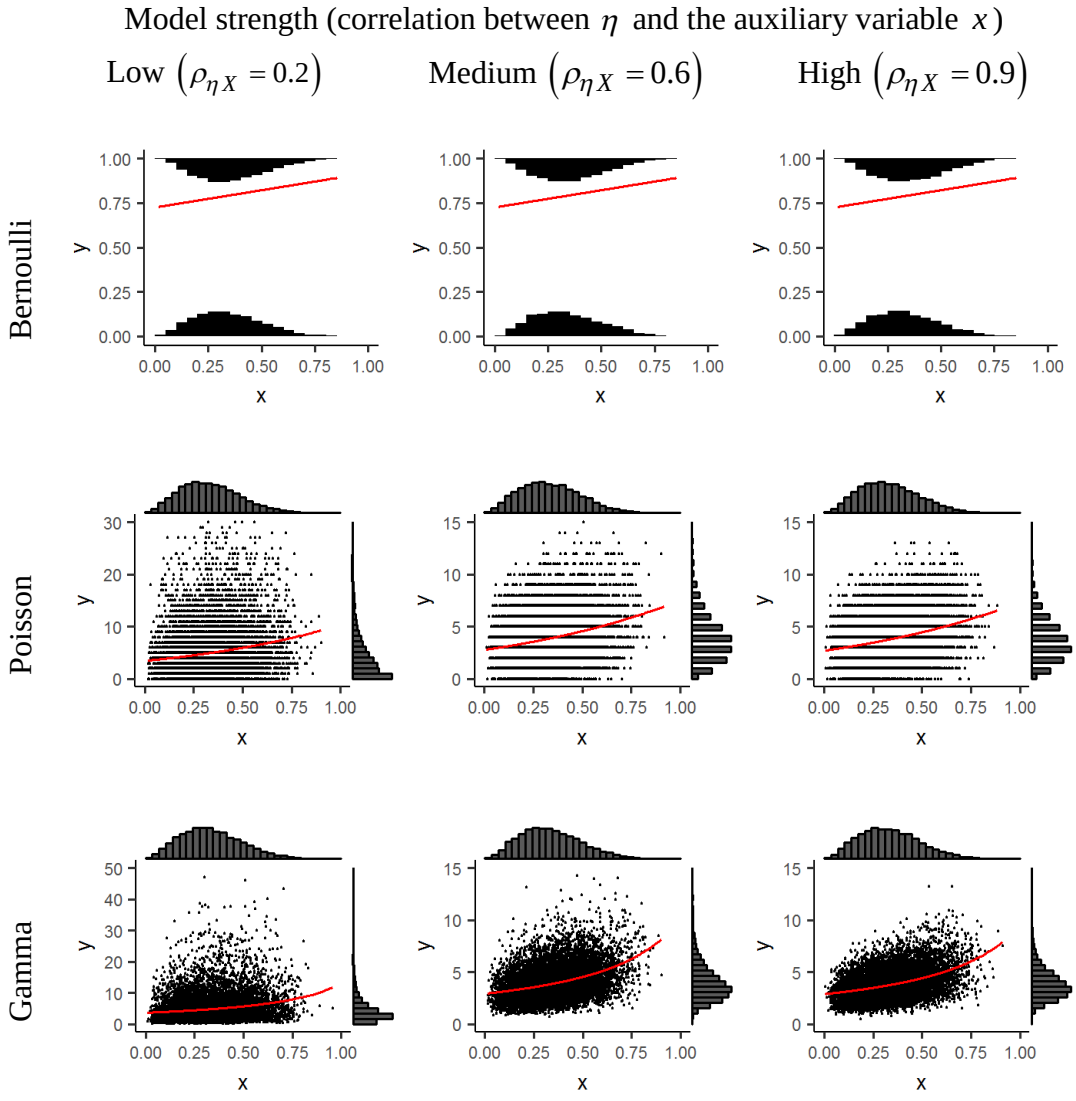
	Population Scenarios/ Distribution of $y_k \mid \mathbf{x}_k$		
	Bernoulli	Poisson	Gamma
Type of outcome	Binary	Count	Positive continuous
Model $g(\mu_k) = \beta_0 + \beta_1 x_k + \sigma \varepsilon_k$			
Parameters of distribution	$p_k = \mu_k$	$\lambda = \mu_k$	$\alpha = 10\mu_k, \beta = 10$
Link function $\eta_k = g(\mu_k)$	$\log(\mu_k / (1-\mu_k))$	$\log(\mu_k)$	$\log(\mu_k)$
Mean $\mu_k = g^{-1}(\eta_k)$	$1/(1+\exp(-\eta_k))$	$\exp(\eta_k)$	$\exp(\eta_k)$
Linear predictor coefficients η_k			
β_0 (intercept)	-1.00	1.00	1.00
β_1 (slope)	10.00	1.00	1.00
σ (high)	7.30	0.73	0.73
σ (medium)	1.99	0.20	0.20
σ (low)	0.72	0.07	0.07
Empirical population statistics*			
Mean $\bar{Y}$	0.86	3.91	5.00
Variance S_Y^2	0.14	4.91	0.02
Mean $\bar{X}$	0.33	0.33	0.33
Variance S_X^2	0.02	0.02	0.02
Correlation $\rho_{\eta X}$	(high) 0.90	(medium) 0.60	(low) 0.20
Correlation ρ_{YX}	0.36	0.27	0.16

* Population statistics are computed as the averages over the 100,000 simulated populations.

In all scenarios, the MC estimators are oracle estimators in the sense that they have a correctly specified mean, although the variance might be misspecified because of the dispersion induced by σ. The linear working models for GREG and LNPA estimators have a misspecified functional form for the type of data. For the nonlinear PA estimators, the functional forms of the working models are correct, but we cannot

say that their working models are correctly specified or misspecified because the variables in the selected model are determined algorithmically.

Figure 2.1 Scatter plots of the populations described in Table 2.5



In all scenarios, the MC estimators are oracle estimators in the sense that they have a correctly specified mean, although the variance might be misspecified because of the dispersion induced by σ. The linear working models for GREG and LNPA estimators have a misspecified functional form for the type of data. For the nonlinear PA estimators, the functional forms of the working models are correct, but we cannot say that their working models are correctly specified or misspecified because the variables in the selected model are determined algorithmically.

In each simulation run, a new population from the sequence of 10 populations with indices $U_N = \{1,...,2000N\}_{N=1}^{10}$ is generated using the model parameters listed in Table 2.5. Each finite population N_N-th of size $N_N \in \{2000, 4000, \ldots, 20000\}$is sampled with $f = 0.05$. For the SRS and PPS samples n_N is fixed, $n_N \in \{100, 200, \ldots, 1000\}$, while for PO design, these are the expected sample sizes. The simulation is run $B = 100{,}000$ times for each scenario, sample design, and population in the sequence. The performance of the seven estimators is evaluated using RB and RE defined in Section A.4 in Appendix A.

The results of the simulations are summarized graphically for the Bernoulli population for the SRS and PO designs in Figures 2.2 and 2.3, respectively. Figure 2.2 shows six plots for the RB for six estimators for the Bernoulli population. In each plot, the vertical axis indicates the RB as a percentage while the horizontal axis is the sample size used to compute the estimator. The first row shows the RB of the estimators computed from samples from an SRS design while the second row is

the RB for samples using a PO design. The columns indicate the values of $\rho_{\eta X}$ which measure the model strength when the population is generated. Figure 2.3 shows plots with the RE with the same layout for the SRS and PO sample designs for the Bernoulli population. The complete set of figures for all populations, sample designs, and models is found in Appendix A, Section A.1.

The first row in Figure 2.2 shows that for the Bernoulli population, the RBs of all estimators are very small, even for samples of 100 cases when using an SRS design. For example, the largest RB is for the MC estimator 0.13% for a sample size of 100 cases for $\rho_{\eta X} = 0.9$. The same pattern holds for all examined populations and correlations for SRS designs.

Although the empirical RBs of the estimators are small, they become noticeable in smaller samples drawn using a PO design as shown in the second row of Figure 2.2. Except for the NLPA estimator, the RBs can be greater than 0.5% for samples of 100 cases for $\rho_{\eta X} = 0.6$ and 0.9. The RBs do not become zero in samples as large as 1,000 cases for the Bernoulli population for $\rho_{\eta X} = 0.9$. The HJ estimator has the largest RB when the correlation is small or medium. In this population, the NLPA estimator has a smaller RB for $\rho_{\eta X} = 0.9$ and approaches to zero for smaller sample sizes when $\rho_{\eta X} = 0.6$ and 0.9. A similar pattern holds this population and the PPS design but with slightly smaller biases; see Figures A.1 through .9 in Section A.1 in Appendix A.

Although not as extreme as in the Bernoulli population, the RBs of the estimators have similar patterns in the Gamma and Poisson populations for the PO and PPS designs with one exception. The NLPA estimator has a considerably larger RB in most of the range of sample sizes examined (i.e., for a sample size of 100 cases, between 1.5 and 2.5 percentage points in the Gamma population and between two and four percentage points in the Poisson population).

We discuss the bias of the HJ estimator, and this discussion applies to other estimators as well. The HJ estimator is a ratio estimator in the PPS and PO designs and its bias, $\mathbb{B}\left(\hat{Y}_{HJ}\right)$, is a function of the covariance between the estimates of $\hat{\bar{Y}}_{HJ}$ and $\hat{N}_{HT}$, $\mathbb{B}\left(\hat{Y}_{HJ}\right) = -\mathbb{C}ov\left(\hat{\bar{Y}}_{HJ}, \hat{N}_{HT}\right)$, where $\hat{N}_{HT} = \sum_{k \in A} d_k$ (see Cochran, 1977). The correlation between y and π in these populations is high by design since both quantities are functions of the auxiliary variable x. Although the bias vanishes in large samples because in the sequence of estimators $\hat{\bar{Y}}_{HJ,N}$ and $\hat{N}_{HT,N}$ are consistent, e.g., $\hat{\bar{Y}}_{HJ,N} - \bar{Y}_N = \mathcal{O}_p\left(n_N^{-1/2}\right)$ and $\hat{N}_{HT,N} - N_N = \mathcal{O}_p\left(n_N^{-1/2}\right)$, the bias is noticeable when n is small in the PPS and PO designs. For example, $\mathbb{B}\left(\hat{Y}_{HJ}\right) = 0.43\ Y\ \%$ for a sample size of 100 cases from the Bernoulli population.

The source of the bias of the NLPA estimator is different since it is not a ratio. As described in Section 1.5.5, the PA adjustment $\hat{\mathbf{\Gamma}}_{\mathbf{X}}$ is applied to the linear predictor $\eta_k = \mathbf{x}_k\boldsymbol{\beta}$. The impact of this adjustment on the estimator depends on the inverse of

the link function, g^{-1}, that maps the PA adjusted η_k to μ_k as $\mu_k = g^{-1}(\eta_k)$. In the Bernoulli population, the inverse of the link function $g^{-1}(\eta_k)$ of the NLPA estimator is the logistic function that bounds $\hat{\mu}_{pa,k}$ to values between zero and one. As a result, the effect of any PA adjustments is controlled, since the PA adjusted mean $\hat{\mu}_{pa,k}$ cannot be greater than one or less than zero.

In contrast, in the Poisson and Gamma populations, the inverse of the link function of the NLPA estimator is the exponential function, $\hat{\mu}_{pa,k} = g^{-1}(\hat{\eta}_{pa,pa}) = \exp(\mathbf{x}_k \hat{\mathbf{\Gamma}}_{\mathbf{X}} \hat{\boldsymbol{\beta}}_{pmle})$, and its support has a lower bound but no finite upper bound (i.e., any positive number greater than zero for the Gamma distribution or greater than or equal to zero for the Poisson distribution). Although the values of the PA adjusted means $\hat{\mu}_{pa,k}$ are stochastic and depend on the ratio of the auxiliary variable population totals and their estimates, the PA estimated mean $\hat{\mu}_{pa,k}$ may be very large after this ratio is exponentiated. As a result, the NLPA estimator is expected to require very large sample sizes to converge. These observations are illustrated in the figures that show small biases for the NLPA estimator at small sample sizes for the Bernoulli population and large biases even with sample sizes as large as 1,000 in the Poisson and Gamma populations.

The RE of the estimators for the SRS design across the populations is almost constant for all the sample sizes in the simulations. In contrast, in the PPS and PO designs, the RE is not constant because the bias component of the MSE differs by sample size. For

this analysis, we use the averages of the RE of groups of estimators across the range of sample sizes to characterize their gains in efficiency in the PPS and PO designs, even though some estimators perform better for specific ranges of sample sizes and populations types.

The first row of Figure 2.3 shows the RE of the estimators for the Bernoulli population for samples drawn using SRS. The RE of the estimators is correlated to the values of $\rho_{\eta X}$. The average RE of the GREG, MC, LNPA, and NLCA estimators are 1.0%, 5.5%, and 7.0% for $\rho_{\eta X} =$ 0.2, 0.6, and 0.9, respectively.

The RE of the estimators of the Bernoulli population for the PO designs is higher than the RE for SRS as shown in the second row of Figure 2.3. If we combined the GREG, MC, LNPA, and NLCA estimators, their average REs are 64.0%, 79.0%, and 83.1% for $\rho_{\eta X} =$ 0.2, 0.6, and 0.9, respectively. When $\rho_{\eta X} =$ 0.6 or 0.9, the estimators with similar values REs form two groups. The first group consists of the GREG and LNPA estimators, and they have a higher RE average than the second group of estimators (the MC and NLCA estimators). The differences in the combined RE average between the first and second group are 4.8% and 7.5% for $\rho_{\eta X} =$ 0.6 and 0.9, respectively. A similar pattern holds for the PPS designs for the Bernoulli population, but with smaller differences in RE.

All estimators are more efficient than the HT estimator for all designs and correlations for the Bernoulli population. The average REs of the HJ and NLPA

estimators are much lower than the average of the other estimators. The MC estimator does not perform as well as the others in the Bernoulli population.

Similar patterns in RE are observed for the Gamma and Poisson populations except for the clustering of estimators with similar RE for high values of $\rho_{\eta X}$. The gains in efficiency from the estimators are generally small in SRS designs compared to the PPS and PO designs, and the gains in the PPS designs are smaller than the gains in the PO designs. In many cases, the linear estimators (GREG and LNPA) have a larger RE than the nonlinear estimators with the correct working model. None of the estimators that use auxiliary information do worse than the estimators that ignore the auxiliary information altogether even when the relationship between the outcome and auxiliary variable is weak.

The estimator with the highest RE varies by scenario and the GREG, LNPA, MC, and NLCA estimators all perform very similarly. The GREG estimator has the largest gain in RE in one-third of the scenarios, followed closely by the LNPA estimator. The performance of the linear estimators is surprising because the linear functional form of the working models is always misspecified for all outcomes. From a practical point of view, none of the four estimators (GREG, LNPA, NLCA, and MC) has a significant advantage over the others across the simulated scenarios for these populations.

The MC estimator, proposed as an estimator that makes more effective use of the auxiliary information from the frame, does no better than the linear estimators (GREG, LNPA) or the nonlinear (NLCA) estimator that only use the population totals

of the auxiliary variables. Furthermore, all the model-assisted estimators and the MC estimator do well even if $\rho_{\eta,x}$ is very low. These results differ from those reported in Wu & Sitter (2001). The MC estimator uses the predicted PML means of the model applied to the whole population; however, it is not clear why this should be more efficient than just using the auxiliary variable population totals. We would not expect substantial gains from the MC estimators in most situations, as shown in these simulations.

A second observation is that the GREG and LNPA estimators perform as well or better than the nonlinear estimators with correctly specified models. This observation questions the reasons for considering nonlinear estimates that are less efficient than their linear counterparts. One answer is that linear estimates, especially for domains, can be negative. Negative estimates are avoided in nonlinear models. This feature is important in nonresponse research, where we use nonlinear models for modeling response propensities.

Figure 2.2 Relative bias (RB) of seven estimators as a function of the sample size for a population with a Bernoulli distribution by sampling design (SRS and PO) by model strength (medium, low, and high).

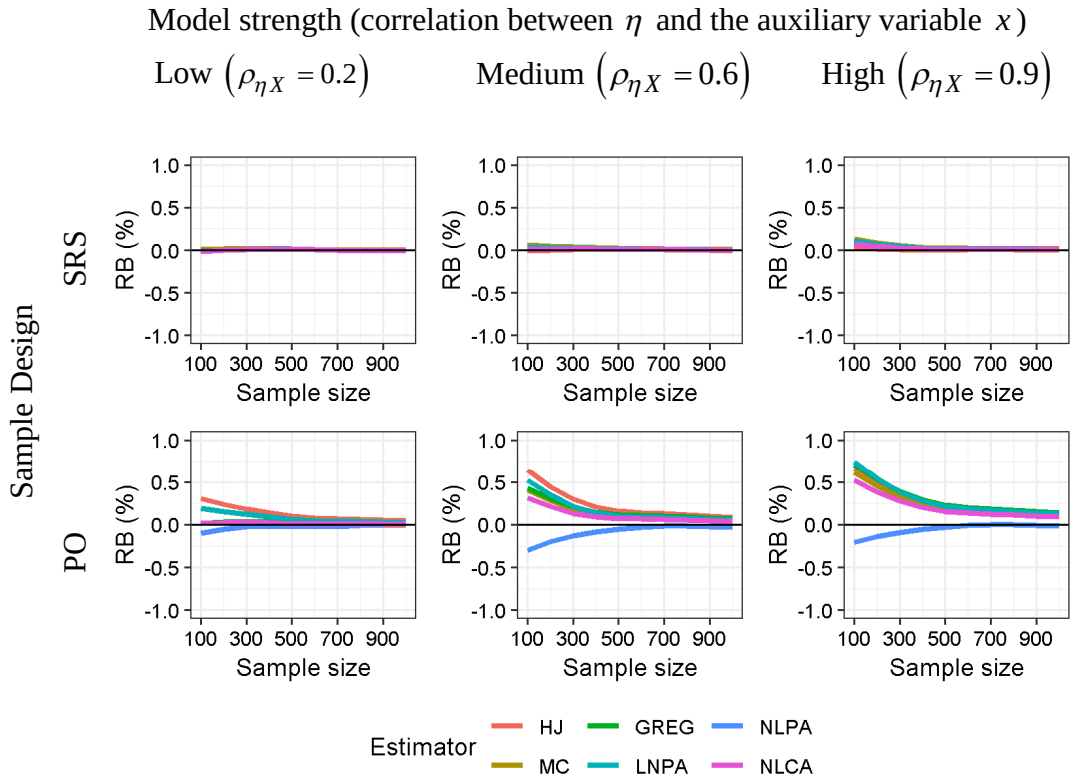
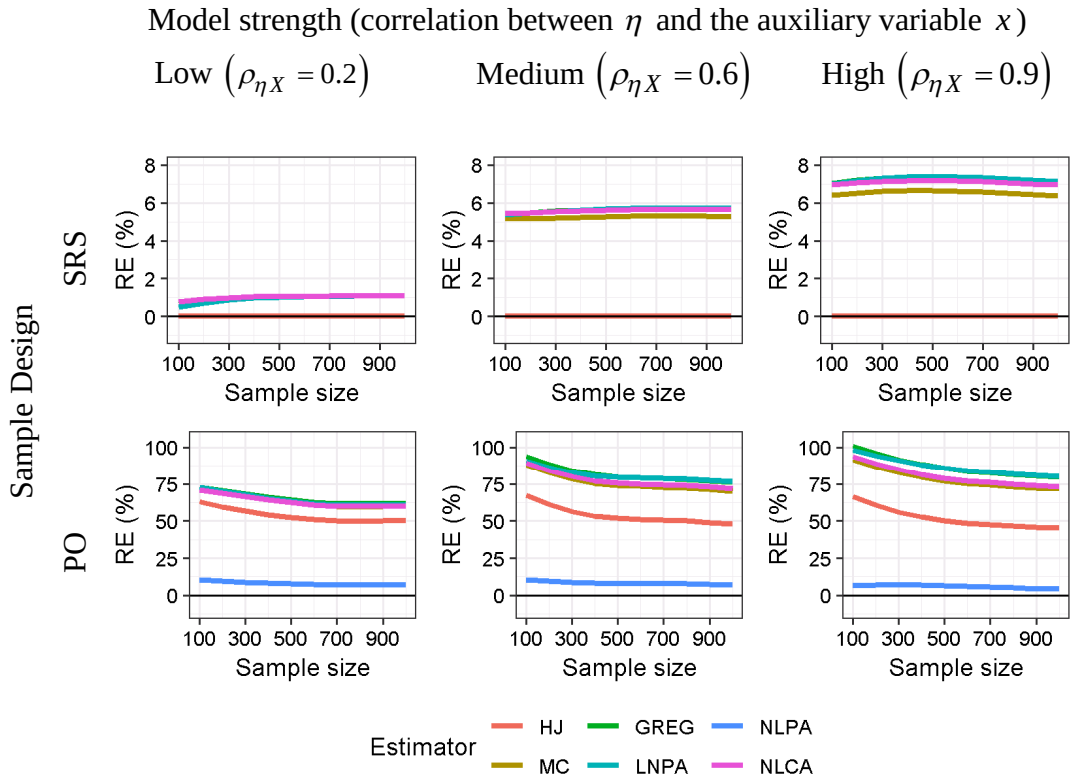


Figure 2.3 Relative efficiency (RE) of seven estimators as a function of the sample size for a population with a Bernoulli distribution by sampling design (SRS and PO) by model strength (medium, low, and high).



2.4 Algorithmic PA Estimators in Poisson Sample Designs

In this example, we derive algorithmic PA estimators for the total $Y = \sum_{k \in U} y_k$ from samples drawn using a Poisson sample design PO, compare these estimators to alternatives found in the literature, and evaluate their statistical properties using simulation.

In a PO sample design, each element has a predetermined positive inclusion probability $\pi_k > 0$ for $k \in U$ (Särndal, Swensson, & Wretman, 1992). Let n_s be the observed sample size (e.g., realized sample), which is a random variable, and n be the expected sample size under repeated sampling defined as $n = \mathbb{E}(n_s \mid \mathcal{F}) = \sum_{k \in U} \pi_k$.

[14]

The PA algorithmic estimators in this example are derived from the working models spanned by the auxiliary variables $\mathbf{x}_k = (1, \pi_k, d_k)$, the unit indicator, the probability of inclusion, and the sampling weight, and their corresponding population totals are $\mathbf{X} = (N, n, d)$ where $d = \sum_{k \in U} d_k$. The estimators considered differ in the complexity of the location and scale parameters of the models. The two algorithmic PA estimators are

---

[14] The PO sample design can be seen as the realized sample of N independent trials, where each element y_k has a probability π_k of appearing in the sample.

1. PA Estimator $\hat{Y}_{PA1}$ with the collection of working models $\mathcal{M}_{1,y}$ with $y \overset{iid}{\sim} \mathcal{N}\left(\theta_\beta, \theta_\sigma^2\right)$ where $\theta_\beta \mid \mathbf{x}_k = \eta_{\beta,k} = \beta_0 + \beta_1 \pi_k + \beta_2 \pi_k^{-1}$ and $\theta_\sigma \mid \mathbf{x}_k = \eta_{\sigma,k} = \sigma_0$, The auxiliary variables for the location parameter are $\left(1, \pi_k, d_k\right)$ with the population totals $\left(N, n, d\right)$. For the scale parameter, the auxiliary variable is 1 with a control total N.

2. PA Estimator $\hat{Y}_{PA2}$ with a collection of more complex working models $\mathcal{M}_{2,y}$ with $y \overset{iid}{\sim} \mathcal{N}\left(\theta_\beta, \theta_\sigma^2\right)$ where

$$\begin{aligned} \theta_\beta &= \mu_k = \beta_0 + \beta_1 \pi_k + \beta_2 d_k, \text{ and} \\ \theta_\sigma &= \left(\exp\left(\sigma_0 + \sigma_1 \pi_k + \sigma_2 d_k\right)\right)\left|\mu_k\right|^{\theta_\gamma / 2}, \end{aligned} \tag{2.4}$$

where $\theta_\gamma = \gamma_0$. The regression models model in (2.5) is more appropriate when the variance is proportional to a power of the mean. The auxiliary variables for the location and scale parameters are $\left(1, \pi_k, d_k\right)$ with the population totals $\left(N, n, d\right)$. For the shape parameter, the auxiliary variable is 1 with a control total N. In the PA estimator $\hat{Y}_{PA2}$, the observed sample determines the working models for the mean and variance.

The simulations below explore the performance of the estimators when the working model is misspecified. The estimators are evaluated for the four scenarios described in Table 2.6 for a population size of $N =$ 10,000 and the expected sample size is $n =$ 500. The superpopulation generating model is

$$y_k = \beta_0 + \beta_1 \pi_k^{\alpha} + \sigma \pi_k^{\gamma} \varepsilon_k, \tag{2.5}$$

where the values of the parameters β_0, β_1, α, σ, and γ are listed in the table, and the error term ε_k is $\mathcal{N}(0,1)$. In all scenarios, a latent or unobserved variable z_k with a distribution $\mathcal{B}eta(3,6)$ is used to compute the measure of size $x_k = 10 + 10\, z_k$, and the first-order inclusion probabilities are $\pi_k = n \dfrac{x_k}{\sum_{k \in U} x_k}$.[15]

In Scenarios 1 and 2, y is positively correlated with π while for Scenarios 3 and 4, this correlation is negative. Scenarios 1 and 3 do not include an intercept term (e.g., $\beta_1 = 0$), while the intercept is nonzero in Scenarios 2 and 4. Since the collection of working models $\mathcal{M}_{1,y}$ assumes a constant variance, the models in $\hat{Y}_{PA1}$ are misspecified in Scenarios 1, 3 and 4. On the other hand, the working models in $\mathcal{M}_{2,y}$ for $\hat{Y}_{PA2}$ can reproduce the correct model for both mean and variance in Scenarios 1, 2 and 3. All working models in the collections $\mathcal{M}_{1,y}$ and $\mathcal{M}_{2,y}$ of the algorithmic PA estimators $\hat{Y}_{PA1}$ and $\hat{Y}_{PA2}$ are misspecified in Scenario 4.

[15] The inclusion probabilities π_k and the auxiliary variable x_k are collinear so either π_k or x_k can be used in the models but not both.

Table 2.6 Parameters of simulations of four scenarios and empirical statistics

Parameters	Model: $y_k = \beta_0 + \beta_1 \pi_k^{\alpha} + \sigma \pi_k^{\gamma} \varepsilon_k$			
	Scenarios			
	1	2	3	4
Population parameters				
β_0	0	10	0	20
β_1	500	500	2	1/8
α	1	1	1	-2
σ	25	5/2	1	6
γ	1/2	0	-1/2	-1/3
Population characteristics				
Empirical population mean $\bar{Y}$	50.00	39.00	40.49	51.86
Empirical population variance S_y^2	62.49	62.49	62.49	62.49
Empirical correlation $\rho_{y\pi}$	0.71	0.74	-0.70	-0.56
Empirical Kish's deff	1.01	1.01	1.01	1.01
Empirical deff the HJ estimator	1.05	1.05	1.10	1.08

In each scenario and simulation run, a new population is generated and sampled using a PO design with an expected sample size $n =$ 500. The estimators of the total $Y = \sum_{k \in U} y_k$ are computed using the realized sample of size n_s. The simulation is repeated $B =$ 100,000 times for each scenario.

The lower pane of Table 2.7 shows selected empirical statistics of the artificial populations such as the mean and variance, the correlation $\rho_{y\pi}$ between $y = \{y_k\}_{k \in U}$ and $\pi = \{\pi_k\}_{k \in U}$, the Kish's weighting design effect,

$deff_{kish} = 1 + \text{cv}(d)^2$ where $d = \{d_k\}_{k \in A}$, and the design effect of $\hat{Y}$ for the HJ estimator $deff_y = \frac{\mathbb{V}\left(\hat{Y}_{HT}\right)}{\mathbb{V}_{SRS}\left(\hat{Y}\right)}$. All these population statistics are computed as the average of statistics of the simulated populations within each scenario. The performance of the estimators is evaluated using RB, RRMSE, and RE with respect to the HT estimator defined in Section A.4 in Appendix A.

The upper pane of Table 2.7 shows the RB, RRMSE, and RE of the HT and HJ estimators used as a reference and the algorithmic PA estimators. The lower panel shows the same statistics for the oracle estimators for each scenario. The oracle estimators are derived as PA estimators assuming there is no model misspecification.(These estimates are algebraic PA estimators and are discussed in Section 1.7.3 on page 85). The results in the table confirm that the HT estimator is very inefficient when the sample is drawn using a PO design, and that the HJ estimator is a better alternative (Särndal, Swensson, & Wretman, 1992). The HJ estimator is on average 95 times more efficient in these scenarios. We are interested in the additional gains in efficiency of algorithmic PA estimators with respect to the HJ and oracle estimators.

We begin the discussion with the empirical bias of the algorithmic PA estimators. As expected for any model-assisted estimators, the RBs are very small in most scenarios.

Now we consider the efficiency of the algorithmic PA estimators $\hat{Y}_{PA1}$ and $\hat{Y}_{PA2}$. In this discussion, we compare the efficiency of $\hat{Y}_{PA1}$ and $\hat{Y}_{PA2}$ to the oracle estimators,

using the oracle for each scenario as a reference except for Scenario 4, where $\hat{Y}_2$ is the reference because there is no oracle. The algorithmic PA estimators are derived using the fit of the model, and they achieve sizeable gains in efficiency over the HJ estimator despite the large initial gains of the HJ over the HT estimator.

The results in Table 2.7 show that, in general, the algorithmic PA estimators track the oracle estimators well even though they do not use the population-generating model. In particular, $\hat{Y}_{PA1}$ with a misspecified and simple model performs as well as the oracle estimators with only a slightly lower RE in Scenarios 1 and 3. These differences are so small that these estimators are practically equivalent.

In Scenario 2, both algorithmic PA estimators are much more efficient than the oracle estimators. They are also more efficient than the best estimator in Scenario 4. The estimator $\hat{Y}_{PA1}$ with a misspecified and simple working model is flexible enough to produce estimates that overcome the negative correlation $\rho_{y\pi}$ that has a large impact on the efficiency of the estimators with a misspecified model. In contrast, the algorithmic PA estimator $\hat{Y}_{PA2}$ with a more complex working model is slightly more efficient than the oracle estimators in Scenarios 1, 3, and 4. An exception is Scenario 2, where $\hat{Y}_{PA2}$ is slightly lower $\hat{Y}_{PA1}$.

The results of the simulations are somewhat surprising. We might expect the algorithmic PA estimator to be much less efficient than the oracle estimators because the algorithmic estimators reflect the increased variance due to the uncertainty of the

model. One hypothesis is that since there are few variables to build the model, the model selection does not contribute significantly to the MSE of algorithmic estimators.

Comparing the two algorithmic estimators, the estimator $\hat{Y}_{PA2}$ with the more complex working model has the largest RE in Scenarios 1, 3, and 4. In contrast $\hat{Y}_{PA1}$ is the best estimator in Scenario 2. However, the differences are very small. These results suggest that using the more complex working model in $\hat{Y}_{PA2}$ gives only small gains in efficiency over $\hat{Y}_{PA1}$. In practice, any of these estimators is a good choice in these scenarios.

These results highlight the importance of a flexible working model, and the exact functional form of the model for the mean and variance is not needed. The results also show that including the inclusion probabilities and the weights as auxiliary variables (if their control totals are available) may improve the efficiency of the estimators. The gains in efficiency and the effect of the model selection with a large number of variables are the topics of future research.

Table 2.7 Empirical relative bias (RB), empirical relative root mean squared error (RRMSE), and empirical relative efficiency (RE) estimator for eight estimators for $n = 500$ and $N = 10{,}000$

Estimator	Scenario*											
	1			2			3			4		
	RB (%)	RRMSE $\times 10^5$	RE	RB (%)	RRMSE $\times 10^5$	RE	RB (%)	RRMSE $\times 10^5$	RE	RB (%)	RRMSE $\times 10^5$	RE
Reference												
HT	0.00	4,382	0	0.01	4,370	0	0.00	4,504	0	0.01	4,804	0
HJ	0.00	687	39.70	0.00	467	86.62	-0.01	691	41.47	-0.01	1,695	7.03
Algorithmic PA												
$\hat{Y}_{PA1}$	0.00	489	79.41	0.00	313	193.81	0.00	496	81.59	-0.02	1,404	10.72
$\hat{Y}_{PA2}$	0.00	489	79.47	0.00	313	193.57	0.00	491	83.06	-0.01	1,393	10.90
Oracle/ Algebraic PA												
$\hat{Y}_1$ (Scenario 1)	0.00	488	**79.59**	0.00	342	162.49	-0.01	1,087	16.18	-0.01	2,007	4.73
$\hat{Y}_2$ (Scenario 2)	0.00	488	79.55	0.00	341	**163.14**	-0.01	1,076	16.52	-0.01	1,994	4.80
$\hat{Y}_3$ (Scenario 3)	0.01	1,082	15.41	0.01	887	23.25	0.00	491	**83.23**	-0.01	1,474	**9.63**

* Scenarios are defined in Table 2.5; RE is the empirical relative efficiency of the estimator with respect to the HT estimator. The empirical estimates are based on 100,000 simulation runs.

Chapter 3 The Algebraic PA Estimators

In Section 1.7, we describe the PA estimators as weighted sums of PA adjusted PML solutions of the working models that relate the outcome y to the auxiliary variables $\mathbf{x}$. This result redefines the role of the working model. In the traditional model-assisted approach, the working model attempts to describe the finite population and leads to a way of estimating model parameters. With the PA, the working model not only guides the functional form of the estimator but is a collection of models that are used to choose the estimator itself and the estimated parameters. This view goes beyond the current understanding of the role of working models in the model-assisted theory.

We can take advantage of this relationship to "engineer" or derive a new class of PA estimators we call *algebraic PA estimators*. To do this, we treat the working model without variable selection. This approach does not utilize a powerful aspect of the PA but does reveal how PA estimators are related to other traditional estimators. In this case, the PA estimator is based on the adjusted pseudo maximum likelihood estimator (PMLE) solution for $\mu_k = \mathbb{E}(y_k)$; if we plug these into the generic form of the PA estimator in Algorithm 3.1, we can produce algebraic PA estimators.

Computing the algebraic PA estimators can be done numerically or algebraically. The latter is often feasible with a linear working model with a few auxiliary variables. In this case, the expression of the algebraic PA estimator may be tractable and can be written in a closed form.

Algorithm 3.1 Algorithm for the derivation of the algebraic PA estimators

Algebraic PA estimators
1: Propose a specific working model $\mathcal{M}_y$ for the outcome y.
2: Compute $\widehat{\mathcal{M}}_y$ with the PMLE of the parameters of the model $\mathcal{M}_y$.
3: Create the PA model $\widehat{\mathcal{M}}_{pa,y}$ by adjusting the PMLE of the regression coefficients $\widehat{\mathcal{M}}_{pmle,y}$ by the PA adjustment $\hat{\mathbf{\Gamma}}_{\mathbf{x}}$.
4: Compute the fitted adjusted PA mean $\hat{\mu}_{pa,k}$ for $k \in A$ using the PA model $\widehat{\mathcal{M}}_{pa,y}$ and substitute $\hat{\mu}_{pa,k}$ in the generic form of the PA estimator $\hat{Y}_{PA} = \sum_{k \in A} d_k \hat{\mu}_{pa,k}$
5: Simplify the expression of $\hat{Y}_{pa} = \sum_{k \in A} d_k \hat{\mu}_{pa,k}$ if it is tractable.

3.1 The Classical Design-Based Estimators as a Class of Algebraic PA Estimators

Some algebraic estimators in the class of linear PA estimators and bias-corrected PA estimators for SRS designs match classical design-based estimators. For example, expansion, stratified, classical ratio, separate ratio, and combined ratio estimators, simple and multiple regression estimators, and poststratified estimator. When the sample design is other than SRS, the PA estimator reproduces generalized versions of these classical design-based estimators. In other words, some classical design-based survey-sampling estimators are a subclass of algebraic PA estimators created using the adjusted PMLE of their working models.

Our rationale for considering algebraic PA estimators is that it provides insights into the conditions when one estimator is more efficient than others. This understanding can inform guidelines for the use of these estimators when there is model uncertainty.

Some prominent estimators are the Hansen, Hurwitz, & Madow (1953) regression estimator, the Hartley & Ross 1954 ratio estimator, the Montanari (1998) randomization optimal estimator, the Deville & Särndal (1992) calibration estimators with a Euclidian distance function, and the Särndal, Swensson, & Wretman (1992) GREG. The list of estimators in the table is by no means complete. For example, the table does not include the alternative design-based estimators for Poisson and Bernoulli sample designs (Särndal, Swensson, & Wretman 1992; Fuller 2009) discussed later.

The view that the classical design-based survey estimators are PA estimators with working regression models with different auxiliary variables has pedagogical value. The PA framework provides a unifying approach to estimation rather than disjoint and seemly unrelated estimators as often presented in sampling textbooks (Cochran 1977; Lohr 2010). However, the PA framework is not fully developed yet, and its current form does not handle complex designs such as multistage sampling.

3.2 Algebraic PA Estimators in Poisson Sample Designs

In this example, we derive three algebraic PA estimators following the steps in Algorithm 3.1 for the total $Y = \sum_{k \in U} y_k$ for samples from a Poisson sample design (PO). The algebraic PA estimators $\hat{Y}_1$, $\hat{Y}_2$, and $\hat{Y}_3$, are evaluated through simulation for four artificial populations generated by the model in (2.5) with population parameters described in Table 2.6 (see Section 2.4. for additional details of these scenarios).

In each scenario, one PA algebraic estimator is an oracle because the estimator is created using the model that generated the population, while the others have a misspecified working model (see Definitions 1.17 and 1.18 in Section 1.5.6). The algebraic PA estimators are:

1. Estimator $\hat{Y}_1$ with a working model for the outcome $y_k \,|\, \pi_k \overset{iid}{\sim} \mathcal{N}\left(\beta_1 \pi_k, \sigma^2 \pi_k\right)$.

   Solving the pseudo-log-likelihood (PL) fitted to the data and simplifying the algebraic expression gives

$$\hat{Y}_1 = \frac{n}{n_s} \hat{Y}_{HT} \,. \tag{3.1}$$

   The expression (3.1) is a ratio estimator where the auxiliary variable is π_k, the estimated total is $n_s = \sum_{k \in A} d_k \pi_k$, and the population total is $n = \sum_{k \in U} \pi_k$. Another way to interpret this estimator is as the Horvitz-Thompson (HT) estimator $\hat{Y}_{HT}$

with a PA adjustment $\hat{\Gamma} = \frac{n}{n_s}$ (see Section 1.7). Särndal, Swensson, & Wretman (1992) propose the estimator in (3.1) as an alternative estimator for Bernoulli (BE) sample designs; however, as shown here, this estimator can also be used in PO sample designs.

If the inclusion probabilities are constant as in BE sample designs, then the PA estimator $\hat{Y}_1$ becomes $\hat{Y}_{alt,1} = N\,\bar{y}_s$ as described in Fuller (1975) and Särndal, Swensson, & Wretman (1992)[16], where $\bar{y}_s$ is the unweighted mean $\bar{y}_s = \sum_{k \in n_s} \frac{y_k}{n_s}$.

The alternative estimator for a PO sample design described by Särndal, Swensson, & Wretman (1992) is $\hat{Y}_{alt2} = N \frac{\hat{Y}_{HT}}{\hat{N}_{HT}}$, which is the Hájek (HJ) ratio estimator of the total Y (Hájek, 1971). The estimator $\hat{Y}_{alt2}$ is itself an algebraic PA estimator with a working model $y_k \,|\, \mathbf{x}_k \overset{iid}{\sim} \mathcal{N}\left(\beta_0, \sigma_0^2\right)$. This PA estimator is the ratio estimator when the auxiliary variable is one instead of π_k. The PA framework justifies the alternative estimators for PO and BE designs proposed in the literature.

[16] Särndal, Swensson, & Wretman (1992) describe a BE as a PO design where the first order probabilities of inclusion are the same, i.e., $\pi_k = \pi$ for $k \in U$.

The algebraic PA estimator $\hat{Y}_1$ in (3.1) is easily generalized to fixed sample designs such as probability proportional to size (PPS) sampling for outcome variables with a working model $y_k \mid \mathbf{x}_k \overset{iid}{\sim} \mathcal{N}\left(\beta_1 x_k, x_k \sigma_1^2\right)$. The auxiliary variable x_k is used as the measure of size for calculating the inclusion probability $\pi_k = n\frac{x_k}{\sum_{k\in U} x_k}$ where $x_k > 0$ for all $k \in U$. Since for this design, $n = n_s$, then $\hat{Y}_1$ reduces to the HT estimator. When this model holds for y_k in PPS sampling, the HT estimator is more efficient than the HJ estimator. This observation identifies one condition where the HT estimator is the preferred estimator. Most discussions in the literature provide arguments in favor of the HJ estimator over the HT estimator, but they do not address the reverse case (Särndal, Swensson, & Wretman, 1992).

2. Estimator $\hat{Y}_2$ with a working model for the outcome $y_k \mid \pi_k \overset{iid}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 \pi_k, \sigma_0^2\right)$. Solving the PL function and simplifying the algebraic expression gives

$$\hat{Y}_2 = \frac{\hat{Y}_{HT}\bar{\pi}_s - n_s\bar{y}_s}{\hat{N}\bar{\pi}_s - n_s} N + \frac{\hat{N}\bar{y}_s - n_s\hat{Y}_{HT}}{\hat{N}\bar{\pi}_s - n_s} n, \tag{3.2}$$

where $\bar{\pi}_s = \frac{\sum_{k\in n_s} \pi_k}{n_s}$ is the unweighted sample mean of the inclusion probabilities of the observed sample n_s. The estimator $\hat{Y}_2$ is the GREG with

auxiliary variables $(1, \pi_k)$, and population totals (N, n) (Särndal, Swensson, & Wretman, 1992).

3. Estimator $\hat{Y}_3$ with a working model $y_k \mid \pi_k \overset{iid}{\sim} \mathcal{N}\left(\beta_1 \pi_k^{-1}, \sigma_1^2 \pi_k^{-1}\right)$. In contrast to previous models, the correlation between the outcome variable and the probability of inclusion is negative. Solving the PL function for this model yields the algebraic PA estimator

$$\hat{Y}_3 = \frac{\widehat{TH}_{\pi,HT}}{TH_\pi} \hat{Y}_{HT}, \tag{3.3}$$

where TH_π is the harmonic total of the inclusion probabilities in the frame, $TH_\pi = N \, \mathrm{H}(\pi)$, where $\mathrm{H}(\pi)$ is the harmonic mean of $\pi = \{\pi_k\}_{k \in U}$ so $\mathrm{H}(\pi) = \frac{N}{\sum_{k \in U} \pi_k^{-1}}$. $\widehat{TH}_{\pi,HT}$ is the HT estimator of TH_π, $\widehat{TH}_{\pi,HT} = \hat{N}_{HT} \, \hat{\mathrm{H}}(\pi)$, where $\hat{\mathrm{H}}(\pi) = \frac{\hat{N}_{HT}}{\sum_{k \in A} d_k \pi_k^{-1}}$. The algebraic PA estimator in (3.3) is a generalization for complex designs of the estimator known as predictive product estimator for SRS proposed by Agarwal & Jain (1989).

The estimator $\hat{Y}_3$ is also a product estimator (Cochran, 1977), and the HT estimator $\hat{Y}_{HT}$ after the PA adjustment $\hat{\Gamma}_{TH} = \left(\frac{TH_\pi}{\widehat{TH}_{\pi,HT}} \right)^{-1}$. As a product estimator, $\hat{Y}_3$ is expected to be more efficient than the ratio estimator (3.1) when

$\{y_k\}_{k\in U}$ is negatively correlated with $\{\pi_k\}_{k\in U}$. The estimator $\hat{Y}_3$ can also be written as the ratio estimator $\hat{Y}_3 = D\frac{\hat{Y}_{HT}}{\hat{d}_{HT}}$, where D is the population total of the weights $D = \sum_{k\in U} d_k$ and $\hat{d}_{HT} = \sum_{k\in n_s} d_k^2$.

The algebraic PA estimators $\hat{Y}_1$, $\hat{Y}_2$, and $\hat{Y}_3$ are oracle estimators for Scenarios 1, 2, and 3, respectively because in these scenarios both the mean and variance of the working models are correctly specified. For Scenario 4 all working models are misspecified. Since the algebraic PA estimators were "engineered" for specific population models (e.g., they are oracle estimators), we focus the discussion on their properties when the models are misspecified.

The lower pane of Table 2.7 shows the relative bias (RB), empirical relative root mean squared error (RRMSE), and relative efficiency (RE) with respect to the HT estimator defined in Section A.4 in Appendix A. The same statistics for the HT and HJ estimators are shown in the upper pane of the table for reference. The highest values of RE are indicated in boldface for each scenario.

We begin by discussing the RB and RE of the algebraic PA estimators. As expected, for any model-assisted estimators, the RBs are very small even if the working models are misspecified (Särndal, 2007). When the model has a good fit, all algebraic PA estimators achieve sizeable gains in efficiency over the HJ, above the substantial gains the HJ has over the HT estimator. The respective oracle estimators have the largest RE in Scenarios 1, 2, and 3.

Estimators $\hat{Y}_1$ and $\hat{Y}_2$ achieve almost the same efficiency in all scenarios. In Scenario 1, the oracle $\hat{Y}_1$ is slightly more efficient than $\hat{Y}_2$. In Scenario 2, the difference between the oracle $\hat{Y}_2$ and $\hat{Y}_1$ is larger but less than a half percentage point. Practically, these differences are very small, and any of these estimators is a good choice in these scenarios. On the other hand, estimators $\hat{Y}_1$ and $\hat{Y}_2$ underperform in Scenarios 3 and 4, where the HJ estimator is two times more efficient in Scenario 3 and five times more efficient in Scenario 4.

The estimator $\hat{Y}_3$ that assumes a negative correlation $\rho_{y\pi}$ is the least efficient estimator in Scenarios 1 and 2 where its working model is grossly misspecified. In these scenarios, the HJ estimator is between 2 and 3 times more efficient than $\hat{Y}_3$. In contrast, $\hat{Y}_3$ is the best estimator in Scenarios 3 and 4 where $\rho_{y\pi}$ is negative. The estimator $\hat{Y}_3$ is between two and five times more efficient than $\hat{Y}_1$ and $\hat{Y}_2$ in Scenario 3 and 4. In Scenario 4, where all working models of the estimator are misspecified, $\hat{Y}_3$ is the best estimator because its working model is closer to the correct model.

These observations highlight the importance of an appropriate working model. We do not need to know the exact functional form of the model for the mean and variance, but the working model should have a reasonable fit. The simulations also show that there are situations when the model-assisted estimator with a grossly misspecified working model can be less efficient than simple estimators such as the HJ estimator. Using the algorithm to choose the models avoids these pitfalls.

Chapter 4 The Theory of the PA Estimators

In this chapter, we describe the theory and motivation of the weighting adjustments of the PA estimator. The weighting procedure is called Orthogonal or Conditional weighting, a procedure initially developed for producing efficient estimators in the presence of nonresponse. Algorithm 1.1 is the result of the modification of the original procedure described in this chapter. The following sections describe the motivation of the PA framework using an analysis based on the statistical concept of propagation of uncertainty (or propagation of errors) in a system. In the last section, we describe extensions of the PA estimator such as estimators with different functional forms and more complex estimators that incorporate additional population characteristics such as the variance, median, and coefficient of variation.

4.1 Orthogonal Weighting

Orthogonal weighting is an analytical methodology for creating weighting adjustments to reduce bias and variance of estimates of survey data. Orthogonal weighting is also called projection weighting since it can be described geometrically as projections of hyperplanes on the vector spaces generated by the span of the auxiliary variables in the models.

We refer to these methods as orthogonal weighting because the auxiliary variables are assumed to be mutually orthogonal or uncorrelated. We discuss departures from the

assumed orthogonality in practice in Section 4. See Chang (2018) for a discussion of orthogonal projection in a related context.

We originally developed this methodology for adjusting for sampling weights for nonresponse; however, we adapted it for the creation of efficient estimators in the presence of full response. The procedure fits parametric models of the outcome variable y, and either the response propensities ϕ when used to adjust for nonresponse, or the probabilities of inclusion π when used for estimation with full response. Although the values of π are known, they are still modeled to identify the auxiliary variables that explain the selection mechanism. To simplify our discussion, we refer to the probability of selection as ϕ in this chapter due to the way the procedure was developed.

The goal of the orthogonal weighting methodology is to identify the smallest set of variables to adjust for nonresponse. For reasons that become apparent later, adjusting using the smallest set of auxiliary variables is the best approach for reducing bias and variance. Orthogonal weighting only targets this group of auxiliary variables related to the probability of response and the survey outcome.

We begin by describing the orthogonal weighting theory as we initially developed it for nonresponse adjustments, followed by the modifications we made so it is applicable for increasing the efficiency of estimates in the presence of full response. Algorithm 1.1 is the result of these modifications.

The principles of Orthogonal Weighting are

1. We fit separate parametric models to the study or outcome variable y and the selection indicator ϕ (either from the selected sample or after nonresponse). When fitting the models, we identify the smallest set of variables for the adjustments. We show that efficient estimators of y can be obtained by adjusting to a smaller set of variables even though there may be a large number of explanatory variables for the study variable y or the probabilities of inclusion ϕ.

2. We do not assume that the true model can be identified (See PA framework Principle 2 on page 29). Misspecified models with omitted variables are possible in the PA approach and are very common in practice (see Definition 1.17 on page 57 for misspecified models).

   While models with extraneous or irrelevant variables do not affect the bias, they can increase the variance of the estimates. Including many extraneous variables in the model reduces the gains in the efficiency of the estimator. The algorithm gives more importance to identifying and excluding extraneous variables when it is used for estimation with full response.

   The views of model misspecification in the orthogonal weighting approach are in sharp contrast with other methodologies that fit complete models under the implicit assumption that more included variables are better than missing any important variables. We show that unbiased and efficient estimators are possible

with orthogonal weighting even if the working models for y and ϕ are both misspecified. This approach is not the current view in the literature especially for double robust estimators (Kim & Haziza, 2014).

3. When adjusting for nonresponse, we require variables that are available for both respondents and nonrespondents. Additional gains are possible if population totals are available for calibration. For estimation in the presence of full response, we require all auxiliary variables to have population totals.

4. We adopt Särndal & Lundström (2005) point of view of the relationship between bias and weighting adjustments. We do not expect the bias of the estimates to be entirely removed by the adjustments, but the bias is mitigated. To reduce bias and increase efficiency, we require powerful auxiliary variables that explain both only the outcome variable(s) and the probabilities of inclusion.

4.2 Effect of Sample Selection in the Distribution of the Observed Data

We examine the effect of the sample selection (either from an informative sample design or from the response mechanism) on the distribution of the outcome variable on the observed sample compared to its distribution in the population. We require an additional assumption; that the outcome variable(s) y and the selection propensities, ϕ, are random variables that can be decomposed as a sum of orthogonal (or uncorrelated) random components. The decomposition of any mechanism into

individual components is a common tool in fields such as engineering, signal processing, physics, mathematics, and measure-theoretic probability. The use of these elementary units does not imply that the data need to conform to this assumption. However, the concept of orthogonal random variables provides a better understanding of the process because we can examine the effect of the procedure on these individual components separately.

We begin by defining the models in terms of orthogonal components. Let $\mathcal{V}_{\mathbf{x}}$ be the P-dimensional space spanned by the vector of auxiliary variables $\mathbf{x} = (x_1, ..., x_P) \in \mathbb{R}^P$, $\mathcal{V}_{\mathbf{x}} = \text{span}(\mathbf{x})$. Since the elements of $\mathbf{x}$ are assumed to be orthogonal among themselves, then $\mathbb{C}(x_p, x_q) = 0$ for all $p \neq q \in \{1, ..., P\}$. We also assume that the vector $\mathbf{x}$ includes all the auxiliary variables of the superpopulation models $\mathcal{M}_y$ for y and $\mathcal{M}_\phi$, for ϕ defined by the linear predictors

$$\begin{aligned} \eta_y &= \mathbf{x}_\beta \boldsymbol{\beta} = \beta_1 x_1 + ... + \beta_{P_\beta} x_{P_\beta}, \text{ and} \\ \eta_\phi &= \mathbf{x}_\phi \boldsymbol{\phi} = \phi_1 x_1 + ... + \phi_{P_\phi} x_{P_\phi}, \end{aligned} \tag{4.1}$$

where $\mathbf{x}_\beta = (x_1, ..., x_{P_\beta}) \in \mathbb{R}^{P_\beta}$ is the vector of orthogonal auxiliary variables associated with the linear predictor η_y, P_β is the dimension of $\mathbf{x}_\beta$ defined as the number of nonzero elements of $\mathbf{x}_\beta$, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_{P_\beta})^{\mathrm{T}} \in \mathbb{R}^{P_\beta \times 1}$ is the vector of

the parameters in η_y.[17] Similarly, $\mathbf{x}_\phi = \left(x_1,...,x_{P_\phi} \right) \in \mathbb{R}^{P_\phi}$ is the vector of orthogonal auxiliary variables associated with the linear predictor η_ϕ, P_ϕ is the number of nonzero elements of $\mathbf{x}_\phi$, and $\phi = \left(\phi_1,...,\phi_{P_\phi} \right)^{\mathrm{T}} \in \mathbb{R}^{P_\phi \times 1}$ is the vector of the parameters in η_ϕ. Notice that the elements in $\mathbf{x}_\beta$ and $\mathbf{x}_\phi$ are not necessarily the same.

Since the random vector $\mathbf{x}$ contains all the auxiliary variables of the models $\mathcal{M}_y$ and $\mathcal{M}_\phi$, then

$$\mathbf{x} = \mathbf{x}_\beta \bigcup \mathbf{x}_\phi = \left(x_1,...,x_P \right) \in \mathbb{R}^P .$$

The vectors, $\mathbf{x}_\beta$ and $\mathbf{x}_\phi$ are subsets of $\mathbf{x}$, e.g., $\mathbf{x}_\beta \subseteq \mathbf{x}$ and $\mathbf{x}_\phi \subseteq \mathbf{x}$.

Let the vector space $\mathcal{V}_\mathbf{x}$ be a subspace of infinite-dimensional vector space $\mathcal{V}_\infty$, where $\mathcal{V}_\infty$ includes other variables that are not part of the models $\mathcal{M}_y$ and $\mathcal{M}_\phi$ but are observed in the sample. The vector subspaces $\mathcal{V}_\beta = \text{span}\left(\mathbf{x}_\beta \right)$ and $\mathcal{V}_\phi = \text{span}\left(\mathbf{x}_\phi \right)$ are both subspaces of $\mathcal{V}_\mathbf{x}$ and since we assume that $\mathcal{V}_\infty$ is an orthogonal space, then this also holds true in $\mathcal{V}_\beta$, and $\mathcal{V}_\phi$.

[17] Without loss of generality and for simplicity, we use only the location parameters. A more formal proof would include the scale and shape parameters of the models.

Since the random vector $\mathbf{x}$ is orthogonal, then $\mathbf{x}$ the basis of the spaces $\mathcal{V}_{\mathbf{x}}$. The subspaces $\mathcal{V}_{\beta}$ and $\mathcal{V}_{\phi}$ are generated by the projection of $\mathbf{x}$ on $\mathbf{x}_{\beta}$ and $\mathbf{x}$ on $\mathbf{x}_{\phi}$, respectively. As a result, $\mathbf{x}_{\beta}$ and $\mathbf{x}_{\phi}$ are the basis of the reduced dimensions of $\mathcal{V}_{\beta}$, and $\mathcal{V}_{\phi}$.

To clarify this setting, consider the superpopulation models $\mathcal{M}_y$ and $\mathcal{M}_{\phi}$ listed in Table 4.1. The table shows the parameters of the models for the outcome variable y and the sample selection ϕ. The first model, $\mathcal{M}_y$, has a linear predictor $\eta_{\beta} = \beta_1 x_{k1} + \beta_3 x_{k3} + \beta_4 x_{k4} + \beta_5 x_{k5}$ with the auxiliary variable vector $\mathbf{x}_{\beta} = (x_1, x_3, x_4, x_5)$. The second model, $\mathcal{M}_{\phi}$, has a linear predictor $\eta_{\phi} = \phi_1 x_1 + \phi_2 x_2 + \phi_3 x_3$ with the auxiliary variable vector $\mathbf{x}_{\phi} = (x_1, x_2, x_3)$. The vector space $\mathcal{V}_{\mathbf{x}}$ that includes all parameters of the models $\mathcal{M}_y$ and $\mathcal{M}_{\phi}$ is spanned by $\mathbf{x} = (x_1, x_3, x_4, x_5) \bigcup (x_1, x_2, x_3) = (x_1, x_2, x_3, x_4, x_5)$. Note that the auxiliary variable x_2 does not play any role in $\mathcal{M}_y$. Similarly, the variables x_4 and x_5 do not play any role in $\mathcal{M}_{\phi}$. Since $\mathbf{x}$ is orthogonal, then $\mathbb{C}(x_p, x_q) = 0$ for $p \neq q \in \{1,..,5\}$.

The assumption of the orthogonal decomposition of random variables with a common base is very strong and is partly justified by the Karhunen-Loève theorem for the

expansion of a stochastic process (Ghanem & Spanos, 2012)[18]. This assumption must be relaxed for orthogonal adjustments for estimation with full response because the auxiliary variables are not orthogonal in practice.

To describe the effect of the sample selection on the distribution of y in the sample, we expand the definitions presented above.

Let $\mathcal{F} = (\mathbf{y}, \mathbf{x}) \in \mathbb{R}^{N \times (P+1)}$ be a finite population generated by N *iid* realizations of the superpopulation model for y, $\mathcal{M}_y$, where $U = \{1, ..., N\}$ are the labels of $\mathcal{F}$, and $\mathbf{x}_k = (x_{k1}, ..., x_{kP}) \in \mathbb{R}^{P \times N}$ is a realization of the random vector $\mathbf{x}$ described above for $k \in U$. Let $\mathbf{y}_N \in \mathbb{R}^{N \times 1}$ be the population vector of the outcome variable y with a distribution function $\mathrm{f}_{Y_k}(y_k) 1_{\{y \in D\}}$ where D is the support of y, and $\mathbb{E}(\mathbf{y}_N) = \mathrm{g}_\beta^{-1}(\boldsymbol{\eta}_\beta) \in \mathbb{R}^{N \times 1}$, $\boldsymbol{\eta}_\beta = \mathbf{x}_\beta \boldsymbol{\beta} \in \mathbb{R}^{N \times 1}$ is the linear estimator, and g_β^{-1} is the inverse of the link function for y.

[18] A set of orthogonal random variables can be obtained from a set or correlated random variables by principal component decomposition or by Gram-Schmidt orthonormalization.

Table 4.1 Example of models for y and ϕ with their associated linear predictors and auxiliary variables

Dependent variable	Model $\mathcal{M}$	Sub-vector space	Linear predictor η	Auxiliary variable vector of the model $\mathcal{M}$	Dimension	Auxiliary Vector* $\mathbf{x}$	Model parameters				
							x_1	x_2	x_3	x_4	x_5
Outcome y	$\mathcal{M}_y$	$\mathcal{V}_\beta$	$\eta_{\beta_k} = \beta_1 x_{k1} + \beta_3 x_{k3} + \beta_4 x_{k4} + \beta_5 x_{k5}$	$\mathbf{x}_\beta = (x_1, x_3, x_4, x_5)$	$P_\beta = 4$	$(x_1, 0, x_3, x_4, x_5)$	β_1	0	β_3	β_4	β_5
Sample selection ϕ	$\mathcal{M}_\phi$	$\mathcal{V}_\phi$	$\eta_{\phi,k} = \phi_1 x_{k1} + \phi_2 x_{k2} + \phi_3 x_{k3}$	$\mathbf{x}_\phi = (x_1, x_2, x_3) =$	$P_\phi = 3$	$(x_1, x_2, x_3, 0)$	ϕ_1	ϕ_2	ϕ_3	0	0

* The dimension of $\mathbf{x}$ is $P = 5$.

Let R_k be the random variable for the indicator for whether the element $k \in U$ is selected in the sample (or is a respondent) or not, defined as

$$R_k = \begin{cases} 1 & \text{if unit } k \text{ is selected in the sample (or responds)} \\ 0 & \text{otherwise} \end{cases}, \tag{4.2}$$

with a probability mass function

$$\mathrm{f}_{R_k}\left(R_k = r_k\right) = \begin{cases} \phi_k & R_k = 1 \\ 1-\phi_k & R_k = 0 \end{cases}. \tag{4.3}$$

The probability of $R_k = 1$ or $R_k = 1$ are functions of ϕ_k. Let $\mathbf{R} = \left[R_k\right] \in \left\{0,1\right\}^{N\times 1}$ be the random vector for the whole population where the population mean vector for $\mathbf{R}$ is $\mathbb{E}\left(\mathbf{R}\right) = \boldsymbol{\phi} = \mathrm{g}_{\phi}^{-1}\left(\boldsymbol{\eta}_{\phi}\right) = \left[\mathrm{g}_{\phi}^{-1}\left(\eta_{\phi k}\right)\right] = \left[\phi_k\right] \in \left(0,1\right)^{N\times 1}$, the linear predictor is $\boldsymbol{\eta}_{\phi} = \mathbf{x}_{\phi}\boldsymbol{\phi} \in \mathbb{R}^{N\times P_{\phi}}$, the inverse of the link function for ϕ is g_{ϕ}^{-1}, and $\mathbf{x}_{\phi} \subseteq \mathbf{x} \in \mathbb{R}^{N\times P_{\phi}}$ is the vector of auxiliary variables of ϕ. The discrete random vector $\mathbf{R} = \left[R_k\right] \in \left\{0,1\right\}^{N\times 1}$ classifies the elements of $\mathcal{F}$ into those that appear in the sample or not depending on the probability of selection $\boldsymbol{\phi} = \mathrm{g}_{\beta}^{-1}\left(\boldsymbol{\eta}_{\beta}\right) \in \left(0,1\right)^{N\times 1}$.

Since we are not interested in the distribution of $\mathbf{R}$ but on the distribution of $\mathbf{y}$ conditioned on the cases in the sample, we define a new random variable for the product of these two random variables.

Let $\mathbf{W}$ be the random vector result of the vector-to-vector valued function $\mathbf{W}:\mathbb{R}^N \to \mathbb{R}^N$, defined as $\mathbf{W}(\mathbf{y},\mathbf{R}) = \mathbf{y} \odot \mathbf{R}$. The probability distribution of $w_k \in \mathbf{W}$, which is the joint distribution of R_k and y_k, is

$$\mathrm{f}_{W_k}(w_k) = f_{Y_k R_k}(y_k, r_k) = \begin{cases} \phi_k \, \mathrm{f}_{Y_k}(y_k) 1_{\{y \in D\}} & R_k = 1 \\ (1-\phi_k) \mathrm{f}_{Y_k}(y_k) 1_{\{y \in D\}} & R_k = 0 \end{cases}. \tag{4.4}$$

The random vector $\mathbf{W} = [W_k] \in \mathbb{R}^{N \times 1}$ corresponds to the outcome cases $\mathbf{y}$ selected in the sample and entries with zero values for those cases not selected in the sample, $\mathbf{y} \odot \mathbf{r}$, where $\mathbf{r} \in \{0,1\}^{N \times 1}$ is the vector with the realizations of $\mathbf{R} = \mathbf{r}$. The conditional distribution function of $\mathbf{W} \mid R_k = 1$, for only the cases observed in the sample, is derived using the definition of conditional distribution function as

$$\mathrm{f}_{W_k | R_k = 1}(w_k) = \frac{\mathrm{f}_{W_k, R_k}(Y_k, R_k = 1)}{\Pr(R_k = 1)} = \mathrm{f}_{Y_k, R_k}(Y_k, R_k = 1). \tag{4.5}$$

Let $\mathbf{x}^*_{\phi k} = \mathbf{x}_{\phi k}\big|_{R_k = 1} = \left(x^*_{k1}, .., x^*_{kP_\phi}\right)$ be the values of $\mathbf{x}_{\phi k}$ when $R_k = 1$ (e.g., x^*_{k1},…, $x^*_{kP_\phi}$ are not random anymore), then (4.5) becomes

$$\mathrm{f}_{W_k | R_k = 1}(w_k) = \mathrm{f}_{Y_k, R_k = 1}\left(Y_k, \mathbf{x}_\phi = \mathbf{x}^*_\phi\right). \tag{4.6}$$

The expression in (4.6) corresponds to the distribution f_{Y_k} of the original model $\mathcal{M}_y$ for y with the linear predictor $\eta_y = \mathbf{x}_\beta \boldsymbol{\beta} = \beta_0 x_0 + ... + \beta_P x_{P_\beta}$ transformed to the distribution f_{Y_k} of a new model $\mathcal{M}^*_y$ with a linear predictor η^*_y containing only the

auxiliary variables in the vector $\mathbf{x}_\beta$ not found in the vector $\mathbf{x}_\phi$. Conditioning on $R_k = 1$ reduces the random space of y.

To clarify this point, consider the models $\mathcal{M}_y$ and $\mathcal{M}_\phi$ listed in Table 4.1. The model $\mathcal{M}_y$ of the outcome y in the population has the linear predictor $\eta_\beta = \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$ while the model for the sample selection $\mathcal{M}_\phi$ has the linear predictor $\eta_\phi = \phi_1 x_1 + \phi_2 x_2 + \phi_3 x_3$. The model of the observed sample $\mathcal{M}_y^*$ of the outcome y has the linear predictor $\eta_\beta^* = \beta_4 x_4 + \beta_5 x_5$, since x_4 and x_5 are the only auxiliary variables in $\mathbf{x}_\beta$ that that are not found in $\mathbf{x}_\phi$. The linear predictor of the observed cases, η_β^*, is a reduced random space because the distribution of y does not depend on the auxiliary variables x_1 and x_2 anymore.

These results have a geometric interpretation. Let $\mathcal{V}_{y\phi}^{\perp}$ be the vector space of the new model $\mathcal{M}_y^*$ (e.g., when we condition $\mathcal{M}_y$ on the cases where $R_k = 1$ for $k \in U$), then $\mathcal{V}_{y\phi}^{\perp}$ is the orthogonal complement of projection of the vector $\vec{\eta}_\beta$ on the vector $\vec{\eta}_\beta$. Returning to the models in Table 4.1, the vector space $\mathcal{V}_{y\phi}$ is the plane spanned by $\text{proj}_{\vec{\eta}_\phi} \vec{\eta}_\beta$, the projection of vector $\vec{\eta}_\beta = \beta_1 \vec{x}_1 + \beta_3 \vec{x}_3 + \beta_4 \vec{x}_4 + \beta_5 \vec{x}_5$ to $\vec{\eta}_\phi = \phi_1 \vec{x}_1 + \phi_2 \vec{x}_2 + \phi_3 \vec{x}_3$, as

$$\text{proj}_{\vec{\eta}_\phi} \vec{\eta}_\beta = \sum_{p=1}^{P} \frac{\vec{\eta}_\beta \bullet \vec{\eta}_{\phi\, p}}{\left\| \vec{\eta}_{\phi\, p} \right\|^2} \vec{\eta}_{\phi\, p} = \beta_1 \phi_1 \vec{x}_1 + \beta_3 \phi_3 \vec{x}_3, \tag{4.7}$$

where $\bullet$ is the dot product and $\vec{x}_p = \left(0,...,x_p,...,0\right)^{\mathrm{T}} \in \mathbb{R}^{P\times 1}$ for $p \in \{1,...,P\}$ is the vector representation of the basis $x_p \in \mathbf{x}$. The orthogonal complement of the subspace $\mathcal{V}_{y\phi}$, $\mathcal{V}_{y\phi}^{\perp}$, represents the reduced random space of $\vec{\eta}_\beta$ that corresponds to the plane $\vec{\eta}_\beta = \beta_4\phi_4\vec{x}_4 + \beta_5\phi_5\vec{x}_5$ orthogonal to the plane $\mathcal{V}_{y\phi}$. The conditioned model $\mathcal{M}_y^*$ for the observed sample, represented by the subspace $\mathcal{V}_{y\phi}^{\perp}$, depends only on the auxiliary variables x_4 and x_5.

The previous observations are key for designing the algorithm for the orthogonal weighting procedure. If we want to adjust for the effect of sample selection imposed by the model $\mathcal{M}_\phi$, we only need to adjust for the auxiliary variables x_1, and x_3 because x_4 and x_5 are not affected by the selection (or response). Hence the name of orthogonal adjustment because we target only those components affected by the sample selection or response mechanism. If we are modeling ϕ (e.g., ϕ is unknown), we do not need to have the correct model $\mathcal{M}_\phi = \left(x_1, x_2, x_3\right)$, since a misspecified (e.g., reduced) model $\mathcal{M}_\phi^* = \left(x_1, x_3\right)$ can restore the population distribution of y.

The expression of the expected value of y in the observed conditioned on the observed case for $k \in U$ is

$$\mathbb{E}\left(\mathbf{W} \mid R_k = 1\right) = \mathrm{g}_\beta^{-1}\left(\boldsymbol{\eta}_\beta^*\right) = \mathrm{g}_\beta^{-1}\left(\mathbf{x}_{\overline{\beta \cap \phi}}\boldsymbol{\beta}\right), \tag{4.8}$$

where $\mathbf{x}_{\overline{\beta \cap \phi}}$ indicates the auxiliary variables in the complement set of the intersection of the elements of the vectors $\mathbf{x}_{\phi}$ and $\mathbf{x}_{\beta}$. A more formal proof of (4.8) requires measure-theoretic probability and advanced linear algebra (Luenberger, 1969; William, 2011).

4.3 Modeling of the Outcome and Sample Selection

The main element of orthogonal weighting is the development of the models for ϕ and y. Separate parametric models are fitted using initial or saturated models with the same set of auxiliary variables for ϕ and y. In this section, we describe the orthogonal weighting adjustment as it was originally developed for estimation for nonresponse.

4.3.1 Modeling the Parameter ϕ

In the first step of an algorithm that adjusts for sample selection based on the orthogonal approach, we fit a parametric model $\widehat{\mathcal{M}}_{\phi}$ to the sample membership indicator (or respondent) in the population or sample. Fitting the model $\widehat{\mathcal{M}}_{\phi}$ is straightforward because we have the indicator $R_k = r_k$ for respondents and nonrespondents for $k \in A$ or cases in the sample or not for $k \in U$. When fitting the model $\widehat{\mathcal{M}}_{\phi}$, the initial model or saturated model should include all variables that explain the selection mechanism independently of the outcome. The goal of the first

step is to produce the best model for the sample selection for all outcome variables. We expect the model fitting procedure (for example, the modeling based on the AIC as the loss function in the PA framework) to identify and remove extraneous variables in the saturated model.

Returning to the models in Table 4.1, the initial model or saturated model for ϕ, $\mathcal{M}_{\phi}$, includes the auxiliary variables $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$, and the selected model by the algorithm is $\widehat{\mathcal{M}}_{\phi} = (x_1, x_2, x_3, 0, 0)$.

4.3.2 Modeling the Outcome Variable y

In the second step, we fit a parametric model $\widehat{\mathcal{M}}_y$ to the outcome variable y. Fitting the model $\widehat{\mathcal{M}}_y$ is more difficult than fitting the model to $R_k = r_k$ because we only observe the selected sample (or respondents), and it may have a different distribution than the population as discussed in Section 4. Our solution is to use the estimate of ϕ, $\hat{\phi}$ from the model $\widehat{\mathcal{M}}_{\phi}$ identified in Section 4.3.1. We use the model $\widehat{\mathcal{M}}_{\phi}$ to produce a sample-selection adjusted set of weights $\hat{d}_k = \frac{1}{\hat{\mathbb{E}}(R_k)} = \frac{1}{\hat{\phi}_k}$ where $\hat{\phi}_k = \mathrm{g}^{-1}(\hat{\eta}_{\phi,k})$, $\hat{\eta}_{\phi,k} = \mathbf{x}_{\phi,k}\hat{\boldsymbol{\phi}}$ and use this new weight when fitting the model $\widehat{\mathcal{M}}_y$. The adjusted weight $\hat{d}_k$ removes the sample-selection bias of y in the sample and restores in expectation the population distribution when the model $\widehat{\mathcal{M}}_{\phi}$ is correct. This result can be expressed as

$$\mathbb{E}\left(\left(\mathbf{y}\odot\mathbf{R}\mid R_k=1\right)\odot\hat{\mathbf{d}}-\mathbf{y}\mid\mathcal{F}\right)=\mathcal{O}\left(\frac{1}{n}\right), \tag{4.9}$$

where $\hat{\mathbf{d}} = \left[\hat{d}_k\right] \in \mathbb{R}^{N\times 1}$ is the vector of the adjusted weights. As in any model fitting procedure, we may not identify the correct model $\mathcal{M}_y$ due to sample variation. For the models shown in Table 4.1, the initial model for y is $\boldsymbol{\mathcal{M}}_y = \left(x_1, x_2, x_3, x_4, x_5\right)$ and the final model identified at this step is $\widehat{\mathcal{M}}_y = \left(x_1, 0, x_3, x_4, x_5\right)$ using the weights derived from the model $\widehat{\mathcal{M}}_\phi = \left(x_1, x_2, x_3, 0, 0\right)$.

4.4 Modeling y Conditioned on the Reduced Model for ϕ

In the third step, we identify a new model for ϕ, $\widehat{\mathcal{M}}_{y\phi}$, with the variables that explain both y and ϕ using the models $\widehat{\mathcal{M}}_\phi$ fitted in Section 4.3.1 and the model $\widehat{\mathcal{M}}_y$, fitted in Section 4.3.2. The new model $\widehat{\mathcal{M}}_{y\phi}$ for ϕ contains the auxiliary variables from the intersection of models $\widehat{\mathcal{M}}_{y\phi} = \widehat{\mathcal{M}}_y \bigcap \widehat{\mathcal{M}}_\phi$. We refer to these variables as the common variables of the models for y and ϕ. The reason for using only the common variables for the reduced model $\widehat{\mathcal{M}}_{y\phi}$ is justified in Section 4.2. Only the common variables are affected by the sample selection and this adjustment targets only these variables.

We then proceed in the same way as described in the previous section. We recompute selection-adjusted weights $\hat{d}_k^*$ using $\hat{\phi}$ from the reduced model $\widehat{\mathcal{M}}_{y\phi}$ as $\hat{d}_k^* = \frac{d_k}{\hat{\phi}_k}$. At this point, we have several options to produce the estimate when there is sample selection bias. Since the focus of this dissertation is estimation in the presence of full response, these options are not discussed here. The extension of the PA framework to estimation with nonresponse will be the topic of a future paper (See Appendix A).

4.5 Developing the PA Algorithm for Estimation with Full Response

The goals of the orthogonal adjustment procedure for estimation with nonresponse described in Section 4.3 differ from when the method is used for estimation with full response. When there is nonresponse, the goal is to remove selection bias. In contrast, when there is full response, the goal is to improve the efficiency of the estimators because there is no selection bias. The modifications made to the procedure described in Section 4.3 change the focus of the orthogonal adjustments from removing bias to increasing efficiency by identifying as many variables related to the outcome as possible. As shown in the next section, the largest improvements in efficiency are achieved when the model includes the variables with the largest contributions to the variance of the model. The following modifications to the procedure described in Section 4.3 are consolidated in Algorithm 1.1.

1. The algorithm fits the model $\widehat{\mathcal{M}}_y$ for the outcome variable y using the sampling weight d_k (Step 3 of Algorithm 1.1). When there is full response, there is no need to use the weight from $\widehat{\mathcal{M}}_\phi$ because y is observed for all sampled cases.

2. The algorithm replaces the adjusted weight $\hat{d}_k = \dfrac{d_k}{\hat{\phi}_k}$ where $\hat{\phi}_k$ is the fitted selection probability from $\widehat{\mathcal{M}}_{\phi y}$ by

$$\hat{d}_k = \frac{d_k}{\hat{\phi}_k} \frac{\sum_{k \in A} d_k}{\sum_{k \in A} d_k \hat{\phi}_k^{-1}}, \tag{4.10}$$

when fitting y the second time (Step 6 of Algorithm 1.1). Since we want to calibrate to as many variables of the model as possible, and the common variables may be the largest contributors to the variance, the adjusted weight in (4.10) increases the likelihood that the common variables will be selected in the final model.

3. The algorithm fits the model $\widehat{\mathcal{M}}_y^*$ for y using the saturated set of auxiliary variables $\mathbf{x}$ (Step 7 of Algorithm 1.1). This is an extra step that refits the model for y accounting for the effect of the common variables.

Algorithm 1.1 is not unique, and several options can be implemented to target the important variables that contribute to the variance of the estimator. One option is to ignore the algorithm, fit a single model, and calibrate using the auxiliary variables in

the final model. This is the procedure used in Nascimiento Silva & Skinner (1997). This option works well for simple random sample designs, but the estimators are not as efficient in small samples and for informative designs when variables related to the outcome are used for sampling.[19]

A second option is to calibrate only to the common variables that explain ϕ and y. This option yields very efficient estimators (on some occasions, estimators that are more efficient than those produced by Algorithm 1.1) when the common variables are large contributors to the variance of the model for y. The concern with this option is that we do not know if the common variables are the largest contributors when fitting the model. When this is not the case, the efficiency is noticeably lower than the estimators from the algorithm.

A third option is to force the common variables into the final model. We separate the common variables from the pool of variables for the model for y. The final estimate is computed by calibrating the common variables and the variables in the final model $\widehat{\mathcal{M}}_y^*$. The resulting estimator is generally efficient, but its efficiency is not as large in small samples. The issue is that this option tends to identify extraneous common variables when the correlation between the probability of selection and the outcome is low.

[19] There are also differences in the method for variable selection between Nascimiento Silva & Skinner (1997) and the PA algorithm. They use *p*-value based stepwise procedures and the mean squared error as the loss function.

Estimators based on Algorithm 1.1 have the best empirical performance among all the options we evaluated. We were surprised that in cases where there is model selection uncertainty, the estimators were slightly better than those estimators with a fixed model based on a complete analysis of the population data (see Sections 1.3 and 2.1).

4.6 The Variance of the Linear PA Estimator as a Function of the Number of Auxiliary Variables in the Model

We explore the variance of PA linear estimators (calibration estimators) as a function of the number of auxiliary variables in their model to determine the strategy to follow when fitting models in the presence of full response. We analyze the variance of estimators using an artificial example under ideal conditions.

EXAMPLE 4.1 Let y be the outcome variable with a superpopulation model $\mathcal{M}_{y,10}$ with $y_k \overset{iid}{\sim} \mathcal{N}\left(\mathbf{x}_k\boldsymbol{\beta},\sigma_y^2\right)$, where $\mathbf{x}_k = \left(x_1,...,x_{10}\right)$ is the vector with 10 auxiliary variables where $x_p \overset{iid}{\sim} \mathcal{N}\left(\mu_x,\sigma_x^2\right)$, $\mu_x = 1$ and $\sigma_x^2 = 3$. for $p \in \{1,...,10\}$, $\boldsymbol{\beta} = \left(\beta_1,...,\beta_{10}\right)^{\mathrm{T}}$ is the vector of the parameters of the model with values $\boldsymbol{\beta} = \left(10,9,8,7,6,5,4,3,2,1\right)^{\mathrm{T}}$, and $\sigma_y^2 = 5^2$. The auxiliary variables are orthogonal random variables, $x_p \perp x_q$; that is, $\mathbb{C}\mathrm{or}\left(x_p, x_q\right) = 0$ for $p \neq q \in \{1,...,10\}$ and $\mathbb{C}\mathrm{or}\left(x_p, x_p\right) = 1$ for $p \in \{1,...,10\}$ (see Section 4.3). Note that a set of orthogonal random variables can be obtained from a set of correlated random variables by

principal component decomposition or by Gram-Schmidt orthonormalization (Arfken, Weber, & Harris, 2015).

Let $\mathcal{F}$ be the finite population consisting of $N =$ 1,000 *iid* realizations from $\mathcal{M}_y$. The elements of $\mathcal{F}$ are identified by the labels $U = \{1,...,1000\}$. A sample A of expected sample size $n =$ 100 is selected according to a Bernoulli sample design with $\pi_k = \frac{n}{N} =$ 0.01. The sample design is defined by the vector $\mathbf{S} \in \{0,1\}^{1,000\times 1}$ with the sample membership indicators with an expected value $\mathbb{E}(\mathbf{S}) = [0.01]^{1,000\times 1}$, the variance-covariance matrix $\mathbb{C}(\mathbf{S}) = \boldsymbol{\Delta}$ where $\Delta_{kk} = \frac{n}{N}\left(1 - \frac{n}{N}\right) = 0.09$ for $k \in U$ and $\Delta_{kl} = 0$ for $k \neq l \in U$. We assume that the population totals $\mathbf{X} = \mathbf{1}^{\mathrm{T}}\mathbf{x} = (X_1,..., X_{10})$ are known. The parameter of interest is the population total $Y = \mathbf{1}^{\mathrm{T}}\mathbf{y}$ where $\mathbf{y} \in \mathbb{R}^{N\times 1}$ and $\mathbf{y} = [y_k]$ for $k \in U$.

In this example, the outcome y is a linear function of 10 auxiliary variables $\mathbf{x}$. We expect the linear PA estimators with working models with close to the complete set of auxiliary variables to have smaller variances than those estimators with smaller sets. We also expect the full PA estimator, the PA estimator with the complete set of auxiliary variables in its model, to have the smallest variance. On the other hand, if no auxiliary variables are used, then the variance of the PA estimator should be the same as the variance of the HT estimator.

To facilitate the notation, let $\hat{Y}_{PA,c}$ be the PA estimator of the total Y where the subscript c indicates the number of auxiliary variables and totals used in the assumed model as indicated in Table 4.2.

Table 4.2 Variance of incomplete PA estimator as a function of the auxiliary variables

PA estimator $\hat{Y}_{PA,c}$	# of auxiliary variables c	Auxiliary variables $\mathbf{x}$	Population totals $\mathbf{X}$	Parameters $\boldsymbol{\beta}$	Notes
$\hat{Y}_{PA,0}$	0	None	None	None	No calibration, $\hat{Y}_{HT}$
$\hat{Y}_{PA,1}$	1	(x_1)	(X_1)	(β_1)	
$\hat{Y}_{PA,2}$	2	(x_1, x_2)	(X_1, X_2)	(β_1, β_2)	
…	…	…	…	…	
$\hat{Y}_{PA,p}$	p	$(x_1,...,x_p)$	$(X_1,...,X_p)$	$(\beta_1,...,\beta_p)$	
…	…	…	…	…	
$\hat{Y}_{PA,9}$	9	$(x_1,...,x_9)$	$(X_1,...,X_9)$	$(\beta_1,...,\beta_9)$	
$\hat{Y}_{PA,10}$	10	$(x_1,...,x_{10})$	$(X_1,...,X_{10})$	$(\beta_1,...,\beta_{10})$	All auxiliary variables in model $\mathcal{M}_y$,

For example, $\hat{Y}_{PA,0} = \sum_{k \in A} d_k y_k$ is the PA estimator with no information while $\hat{Y}_{PA,10} = \mathbf{X}\hat{\boldsymbol{\beta}}_{pmle}$ is the full PA estimator with an assumed working model $\mathcal{M}_{10,y}$ with the vector of auxiliary variables $\mathbf{x} = (x_1,...,x_{10})$.

Using the result from (1.36) and the definition of an incomplete PA estimator (Definition 1.19), the expression of $\hat{Y}_{PA,0}$ is

$$\hat{Y}_{PA,0} = \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}_{pmle}. \tag{4.11}$$

To compute the variance of $\hat{Y}_{PA,0}$, we note that the assumed model is a valid PA model; therefore the sum of the weighted residuals, $\sum_{k\in A} d_k e_k = 0$, where $e_k = y_k - \mathbf{x}_k\hat{\boldsymbol{\beta}}_{pmle}$. Since the HT estimator is $\hat{Y}_{HT} = \sum_{k\in A} d_k x_k$, then we can rewrite $\hat{Y}_{PA}$ as

$$\hat{Y}_{PA,0} = \sum_{k\in A} d_k \mathbf{x}_k\hat{\boldsymbol{\beta}}_{pmle} = \hat{Y}_{HT}. \tag{4.12}$$

As a result, $\mathbb{V}\left(\hat{Y}_{PA,0} \mid \mathcal{F}\right) = \mathbb{V}\left(\hat{Y}_{HT}\right)$ (see Section 1.7.6). For the sample design in this example, the variance $\mathbb{V}\left(\hat{Y}_{HT}\right)$ is

$$\mathbb{V}\left(\hat{Y}_{HT} \mid \mathcal{F}\right) = \mathbb{V}\left(\hat{Y}_{PA,0} \mid \mathcal{F}\right) = (d-1)\sum_{k\in U} y_k^2 = 3{,}741{,}156, \tag{4.13}$$

where $d = \frac{1}{\pi} =$ 1000 is the sampling weight. Fitting any model without using the population totals does not improve the variance of the PA estimator over the HT estimator.

We now compute the variance of the PA estimator $\hat{Y}_{PA,1}$ with an assumed working model $\mathcal{M}_{1,y}$ with $y_k \overset{iid}{\sim} \mathcal{N}\left(\beta_1 x_{k1}, \sigma^2\right)$; that is, the model with the first auxiliary

variable x_1. The expression of $\hat{Y}_{PA,1}$, the partial PA estimator with one auxiliary variable x_1 and the population total X_1, is

$$\hat{Y}_{PA,1} = X_1 \hat{\beta}_{1,\,pmle}\,. \tag{4.14}$$

If we assume large samples in this example so the effect of the g-factors is not important (see Section 1.7.4), then the variance of $\hat{Y}_{PA,1}$ is

$$\mathbb{V}\left(\hat{Y}_{PA,1} \mid \mathcal{F}\right) \approx (d-1) \sum_{k \in U} e_{1,k}^2 = 2{,}843{,}377\,, \tag{4.15}$$

where $e_{k1} = y_k - x_{k1}\hat{\beta}_{mle,1}$ is the residual of the model $\mathcal{M}_{1,y}$ fitted to the population for $k \in U$. The reduction of variance between $\mathbb{V}\left(\hat{Y}_{HT} \mid \mathcal{F}\right)$ in (4.13) and $\mathbb{V}\left(\hat{Y}_{PA,1} \mid \mathcal{F}\right)$ in (4.15) is 897,778 or 24 percent.

If we assume that the working model is $\mathcal{M}_{2,y}$ with $y_k \overset{iid}{\sim} \mathcal{N}\left(\beta_1 x_{k1} + \beta_2 x_{k2}, \sigma^2\right)$; that is, a working model with the auxiliary variables x_1 and x_2, then the variance of $\hat{Y}_{PA,2}$ of the PA estimator $\hat{Y}_{PA,2}$ is

$$\mathbb{V}\left(\hat{Y}_{PA,2} \mid \mathcal{F}\right) \approx (d-1) \sum_{k \in U} e_{2,k}^2 = 2{,}110{,}065\,, \tag{4.16}$$

where $e_{k2} = y_k - \left(x_{k1}\hat{\beta}_{mle,1} + x_{k2}\hat{\beta}_{mle,2}\right)$ for $k \in U$. The reduction of variance between $\mathbb{V}\left(\hat{Y}_{PA,2} \mid \mathcal{F}\right)$ and $\mathbb{V}\left(\hat{Y}_{PA,1} \mid \mathcal{F}\right)$ is 733,313 (26 percent). The reduction of variance between $\mathbb{V}\left(\hat{Y}_{PA,2} \mid \mathcal{F}\right)$ and $\mathbb{V}\left(\hat{Y}_{HT} \mid \mathcal{F}\right)$ is 1,631,091 (44 percent).

Table 4.3 shows the variances $\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right)$ for the estimator $\hat{Y}_{PA,c}$ for $c \in \{0,...,10\}$ computed as described above. The table also shows the values of variance reduction and percentages with respect to $\hat{Y}_{HT} = \hat{Y}_{PA,0}$ and $\hat{Y}_{PA,c-1}$.

Table 4.3 Variance of incomplete PA estimator as a function of the auxiliary variables

PA estimator $\hat{Y}_{PA,c}$	# of auxiliary variables (c)	Variance	Variance reduction with respect to $\hat{Y}_{PA,c-1}$		Variance reduction with respect to $\hat{Y}_{PA,0}$	
			Value	(%)	Value	(%)
$\hat{Y}_{PA,0}$	0	3,741,156	NA	NA	NA	
$\hat{Y}_{PA,1}$	1	2,843,378	897,778	24	897,778	24
$\hat{Y}_{PA,2}$	2	2,110,065	733,313	26	1,631,091	44
$\hat{Y}_{PA,3}$	3	1,526,610	583,455	28	2,214,546	59
$\hat{Y}_{PA,4}$	4	1,089,527	437,084	29	2,651,630	71
$\hat{Y}_{PA,5}$	5	749,149	340,377	31	2,992,007	80
$\hat{Y}_{PA,6}$	6	527,259	221,890	30	3,213,897	86
$\hat{Y}_{PA,7}$	7	363,427	163,833	31	3,377,730	90
$\hat{Y}_{PA,8}$	8	279,133	84,294	23	3,462,024	93
$\hat{Y}_{PA,9}$	9	248,846	30,287	11	3,492,311	93
$\hat{Y}_{PA,10}$	10	238,349	10,496	4	3,502,807	94

The last row of Table 4.3 shows that the PA estimator $\hat{Y}_{PA,10}$, which uses the correct working model $\mathcal{M}_{y,10}$, achieves the lowest variance with a reduction of 94 percent

with respect to the variance of the PA estimator $\hat{Y}_{PA,0}$ with no auxiliary information. The table also shows that the reduction of variance for this example is not constant for each variable added to the working model. The largest and smallest reduction of variance are achieved when the auxiliary variables x_1 and x_{10} with associated regression coefficients $\beta_1 = 10$ and $\beta_{10} = 1$ are included in the working model, respectively. These results suggest that the strategy for the development of the working model in the presence of full response should target all variables of the true model and not a subset (for example, using only the common variables described in Section 4.4). Although the common variables may be the auxiliary variables with the largest reduction of variance, we do not know if this is the case when fitting the model.

We can derive the algebraic expression for the empirical results presented in Table 4.3 by rewriting the variance $\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right)$ in terms of the variance of the $\mathbb{V}\left(\hat{Y}_{PA,P} \mid \mathcal{F}\right)$, the variance of the PA estimator $\hat{Y}_{PA,P}$ with the full model (e.g., P auxiliary variables). First, we generalize the expression (4.16) so the variance $\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right)$ for the PA estimator $\hat{Y}_{PA,c}$ with c auxiliary variables $\mathbf{x}_c = \left(x_1, ..., x_c\right)$ is

$$\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right) \approx (d-1)\mathbf{e}_c^{\mathrm{T}}\mathbf{e}_c, \tag{4.17}$$

where $\mathbf{e}_c = \mathbf{y} - \mathbf{x}_c\hat{\boldsymbol{\beta}}_{mle,c}$ and $\hat{\boldsymbol{\beta}}_{mle,c} = \left(\hat{\beta}_{mle,1}, ..., \hat{\beta}_{mle,c}\right)^{\mathrm{T}}$. The expression of the difference of the variance of the PA estimator $\hat{Y}_{PA,c}$ with a working model with

c auxiliary variables and PA estimator $\hat{Y}_{PA,P}$ with the full model (or the P auxiliary variables) after algebraic simplification is

$$\begin{aligned}\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right) - \mathbb{V}\left(\hat{Y}_{PA,P} \mid \mathcal{F}\right) &\approx (d-1)\left\{\mathbf{e}_c^{\mathrm{T}}\mathbf{e}_c - \mathbf{e}_P^{\mathrm{T}}\mathbf{e}_P\right\} \\ &\approx (d-1)\hat{\boldsymbol{\beta}}_{mle,c+1}^{\mathrm{T}}\mathbf{x}_{c+1}^{\mathrm{T}}\mathbf{x}_{c+1}\hat{\boldsymbol{\beta}}_{mle,c+1}\end{aligned}, \tag{4.18}$$

where $\mathbf{x}_{c+1} = (x_{c+1},...,x_P)$ and $\hat{\boldsymbol{\beta}}_{mle,c+1} = \left(\hat{\beta}_{mle,c+1},...,\hat{\beta}_{mle,P}\right)^{\mathrm{T}}$ for $c \in \{0,...,P-1\}$.

Notice that $\mathcal{Q}_{\mathbf{x}_c}\left(\hat{\boldsymbol{\beta}}_{mle,c}\right)$, the quadratic form of $\hat{\boldsymbol{\beta}}_{mle,c}$ and the matrix $\mathbf{x}_c^{\mathrm{T}}\mathbf{x}_c$, is $\hat{\boldsymbol{\beta}}_{mle,c}^{\mathrm{T}}\mathbf{x}_c^{\mathrm{T}}\mathbf{x}_c\hat{\boldsymbol{\beta}}_{mle,c}$. Since the matrix $\mathbf{x}_c^{\mathrm{T}}\mathbf{x}_c$ is positive semidefinite, then $\mathcal{Q}_{\mathbf{x}_c}\left(\hat{\boldsymbol{\beta}}_{mle,c}\right) \geq 0$ for any $c \in \{1,...,P\}$. This result shows that the variance of the full PA estimator $\hat{Y}_{PA,P}$ is always equal to or smaller than the variance of the partial PA estimators $\hat{Y}_{PA,c}$ (e.g., $\mathbb{V}\left(\hat{Y}_{PA,P}\right)$ is a lower bound). We can rewrite (4.18) using the lower bound of $\mathcal{Q}_{\mathbf{x}_c}\left(\hat{\boldsymbol{\beta}}_{mle,c}\right)$ as

$$\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right) \leq \mathbb{V}\left(\hat{Y}_{PA,P} \mid \mathcal{F}\right) + (d-1)\lambda_{\max\left(\mathbf{x}_{c+1}^T\mathbf{x}_{c+1}\right)}\left\|\hat{\boldsymbol{\beta}}_{mle,c+1}\right\|_2^2, \tag{4.19}$$

where $\lambda_{\max\left(\mathbf{x}_{c+1}^T\mathbf{x}_{c+1}\right)}$ is the largest eigenvalue of the matrix $\mathbf{x}_{c+1}^{\mathrm{T}}\mathbf{x}_{c+1}$ and $\left\|\hat{\boldsymbol{\beta}}_{mle,c+1}\right\|_2^2$ is the squared $L-2$ norm of the vector $\hat{\boldsymbol{\beta}}_{mle,c+1}$ computed as $\left\|\hat{\boldsymbol{\beta}}_{mle,c+1}\right\|_2^2 = \hat{\boldsymbol{\beta}}_{mle,c+1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{mle,c+1} = \sum_{k \in \{c+1,...,P\}} \hat{\beta}_{mle,k}^2$. Note that in Example 4.1, the eigenvalues of $\mathbf{x}^{\mathbf{T}}\mathbf{x}$ have the same value; that is, $\lambda = \lambda_p = 1{,}000$ for $p \in \{1,...,10\}$.

Furthermore, any of the c eigenvalues of the submatrix $\mathbf{x}_c^{\mathrm{T}}\mathbf{x}_c$ formed by any subvector of auxiliary variables $\mathbf{x}_c \subset \mathbf{x}$ also have the same values $\lambda = 1{,}000$. This is due to the orthogonality of the vector $\mathbf{x}$. Using these results, we rewrite (4.19) as

$$\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right) = \mathbb{V}\left(\hat{Y}_{PA,P} \mid \mathcal{F}\right) + (d-1)\lambda \left\|\hat{\boldsymbol{\beta}}_{mle,c+1}\right\|_2^2, \tag{4.20}$$

Since the eigenvalues are the same for all models in this example, the reduction of variance when adding auxiliary variables to the working model is a function of $\left\|\hat{\boldsymbol{\beta}}_{mle,c}\right\|_2^2$.

Table 4.4 shows the algebraic expression of the variances of the sequence of partial PA estimators $\hat{Y}_{PA,c}$ for $c \in \{0,...,P\}$ using (4.20). The table shows the variance of the incomplete PA estimator decreases as more auxiliary variables are used until the incomplete PA estimator becomes the complete estimator $\hat{Y}_{PA,P} = \hat{Y}_{PA}$ with the lowest variance. The second term $(d-1)\lambda\left(\hat{\beta}_{mle,1}^2 + \hat{\beta}_{mle,2}^2 + ... + \hat{\beta}_{mle,P-1}^2 + \hat{\beta}_{mle,P}^2\right)$ decreases as each auxiliary variable x_p for $p \in \{1,...,P\}$ is added to the working model until it becomes zero. Since the differences of the variance between two consecutive partial PA estimators $\hat{Y}_{PA,c-1}$ and $\hat{Y}_{PA,c}$ are always positive (e.g., $\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right) - \mathbb{V}\left(\hat{Y}_{PA,c-1} \mid \mathcal{F}\right) = \lambda\beta_{mle,c}^2$), the minimum variance is achieved when the estimator is the complete calibration estimator $\hat{Y}_{PA}$.

Table 4.4 Variance of partial PA estimators as a function of the number of auxiliary variables in their model

Estimator		Number of auxiliary variables c	Variance $\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right)$
Partial	$\hat{Y}_{PA,0}$ or $\hat{Y}_{HT}$	0 (No calibration)	$\mathbb{V}\left(\hat{Y}_{PA,P}\right)+(d-1)\lambda\left(\hat{\beta}^2_{mle,1}+\hat{\beta}^2_{mle,2}+\ldots+\hat{\beta}^2_{mle,P-1}+\hat{\beta}^2_{mle,P}\right)$
	$\hat{Y}_{PA,1}$	1	$\mathbb{V}\left(\hat{Y}_{PA,P}\right)+(d-1)\lambda\left(\hat{\beta}^2_{mle,2}+\ldots+\hat{\beta}^2_{mle,P-1}+\hat{\beta}^2_{mle,P}\right)$
	$\hat{Y}_{PA,2}$	2	$\mathbb{V}\left(\hat{Y}_{PA,P}\right)+(d-1)\lambda\left(\hat{\beta}^2_{mle,3}+\ldots+\hat{\beta}^2_{mle,P-1}+\hat{\beta}^2_{mle,P}\right)$
	…	…	…
	$\hat{Y}_{PA,P-2}$	P-2	$\mathbb{V}\left(\hat{Y}_{PA,P}\right)+(d-1)\lambda\left(\hat{\beta}^2_{mle,P-1}+\hat{\beta}^2_{mle,P}\right)$
	$\hat{Y}_{PA,P-1}$	P-1	$\mathbb{V}\left(\hat{Y}_{PA,P}\right)+(d-1)\lambda\hat{\beta}^2_{mle,P}$
Full	$\hat{Y}_{PA,P}$ or $\hat{Y}_{PA}$	P (calibrated to all variables)	$\mathbb{V}\left(\hat{Y}_{PA,P}\right)$

EXAMPLE 4.2 We now examine the effect of including extraneous variables in the working model using the population and sample design from Example 4.1. We assume that there are an additional 10 orthogonal extraneous variables $\mathbf{x} = (x_{11},...,x_{20})$ with $x_p \overset{iid}{\sim} \mathcal{N}(0,3)$ for $p \in \{11,...,20\}$, where $x_p \perp x_q$ for $p \neq q \in \{1,...,20\}$.

Table 4.5 shows the variance $\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right)$ of the sequence of the PA estimators $\hat{Y}_{PA,c}$ for $c \in \{10,...,20\}$ beginning with the correct working model $\widehat{\mathcal{M}}_{y,10}$ with the auxiliary variables $(x_1,...,x_{10})$ after adding the extraneous variables $(x_{11},...,x_{20})$ one at the time to the model $\widehat{\mathcal{M}}_{y,10}$. The table also shows the value of variance reduction and percentages with respect to $\mathbb{V}\left(\hat{Y}_{PA,10} \mid \mathcal{F}\right)$, the variance of $\hat{Y}_{PA,10}$ with the correct model, and $\mathbb{V}\left(\hat{Y}_{PA,0} \mid \mathcal{F}\right)$ to the variance of $\hat{Y}_{PA,0} = \hat{Y}_{HT}$ with no auxiliary variables.

Table 4.5 Variance of incomplete PA estimator as a function of the extraneous variables

PA estimator $\hat{Y}_{PA,c}$	# of auxiliary variables (c)	Variance	Variance reduction with respect to $\hat{Y}_{PA,10}$		Variance reduction with respect to $\hat{Y}_{PA,0}$	
			Value	(%)	Value	(%)
$\hat{Y}_{PA,10}$	10	238,349	NA	NA	3,502,807	93.6
$\hat{Y}_{PA,11}$	11	238,342	897,778	24	3,502,815	93.6
$\hat{Y}_{PA,12}$	12	237,606	733,313	26	3,503,551	93.6
$\hat{Y}_{PA,13}$	13	237,581	583,455	28	3,503,576	93.6
$\hat{Y}_{PA,14}$	14	237,499	437,084	29	3,503,658	93.7
$\hat{Y}_{PA,15}$	15	237,366	340,377	31	3,503,791	93.7
$\hat{Y}_{PA,16}$	16	236,877	221,890	30	3,504,279	93.7
$\hat{Y}_{PA,17}$	17	236,868	163,833	31	3,504,289	93.7
$\hat{Y}_{PA,18}$	18	236,858	84,294	23	3,504,298	93.7
$\hat{Y}_{PA,19}$	19	236,752	30,287	11	3,504,404	93.7
$\hat{Y}_{PA,20}$	20	236,457	10,496	4	3,504,699	93.7

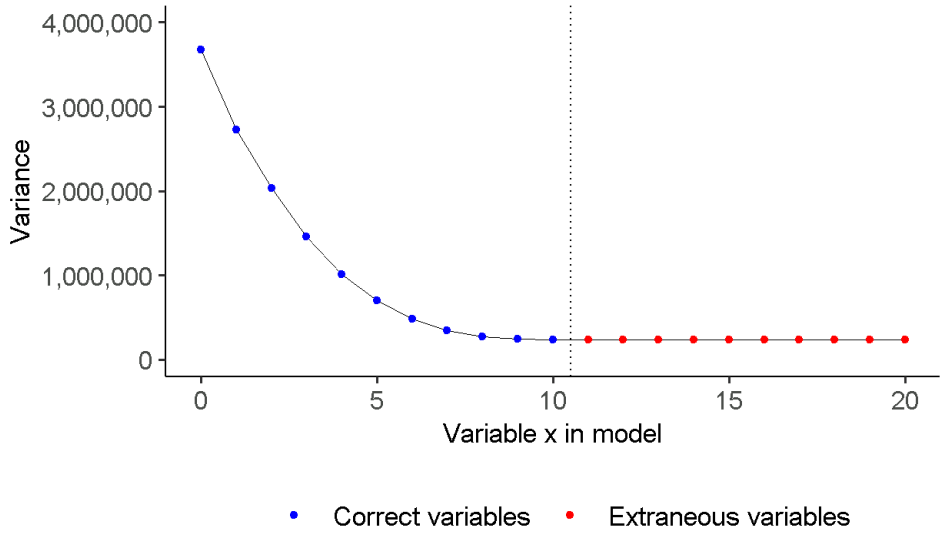
When we fit a variable that is not part of the model, the fitted value of the associated regression coefficient of this variable is zero. As a result, the extraneous auxiliary variables do not contribute significantly to the sum of the squared residuals of the estimator. The expression of the variance of $\hat{Y}_{PA,c}$, as a function of the cumulative number of extraneous variables, is

$$\mathbb{V}\left(\hat{Y}_{PA,c} \mid \mathcal{F}\right) = \mathbb{V}\left(\hat{Y}_{PA} \mid \mathcal{F}\right) + \left(d - 1\right)\mathbf{e}_c^{\mathrm{T}}\mathbf{e}_c\,, \tag{4.21}$$

where $\mathbf{e}_c = \mathbf{y} - \mathbf{x}_c\hat{\boldsymbol{\beta}}_{mle,c}$ for $c \in \{11,...,20\}$. Note that under ideal conditions such as orthogonal variables and very large sample sizes, calibrating to the extraneous variables does not increase the variance of the PA estimator. Based on this analysis, an algorithm should calibrate to as many auxiliary variables as possible to achieve the lowest value of the variance even if it includes extraneous variables as these do not increase the variance of the estimators under these conditions.

Figure 4.1 summarizes graphically the variance reduction from Examples 4.1 and 4.2 for a sequence of PA estimator $\hat{Y}_{PA,c}$ for $c \in \{1,...,20\}$. The line in red show the variance of the PA estimator with a working model where one auxiliary variable is added the time until the complete model (correct) is fitted (e.g., $(x_1,...,x_{10})$ with $\boldsymbol{\beta} = (10,...,1)$). The line in blue shows the variance of the PA estimators beginning with the correct working model when one extraneous variable is added at the time to the correct model for $(x_{11},...,x_{20})$. As shown above, the largest reduction in variance is when the auxiliary variable x_1 with $\beta_1 = 10$ is fitted. Note that although the auxiliary variables x_8, x_9, and x_{10} with associated regression coefficients $\beta_8 = 3$, $\beta_9 = 2$, and $\beta_{10} = 1$ are part of the true model, they do not significantly reduce the variance of the PA estimator.

Figure 4.1 Variance reduction of the sequence of PA estimators from Examples 4.1 and 4.2



4.7 The Propagation of Error for Variance Reduction

Propagation of uncertainty or error is a statistical method that examines how the errors of variables are transmitted through a function in a system (Clifford, 1973). Controlling the propagation of uncertainty is done through adjustments to the input of the functions, so the uncertainty of the function is reduced. We illustrate how the analysis of propagation of errors can provide a better understanding of estimators when they are analyzed as functions of random variables.

Let $\mathbf{S} = [S_k] \in \{0,1\}^{N\times 1}$ be the discrete random vector with the sample membership indicators for a fixed sample. The vector $\mathbf{S}$ follows a discrete multinomial

distribution with $\mathbb{E}(\mathbf{S}) = \boldsymbol{\pi} \in (0,1)^{N\times 1}$ and $\mathbb{C}(\mathbf{S}) = \boldsymbol{\Delta} \in \mathbb{R}^{N\times N}$ (see Definition 1.5). The vector $\mathbf{S}$ is a vector field with $\{0,1\}^N$ where each S_k is a vector.

Define f as the vector-to-scalar valued function $\mathrm{f}: \mathbb{R}^N \to \mathbb{R}$ as

$$\mathrm{f}(\mathbf{S}) = \mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{S}). \tag{4.22}$$

The equation (4.22) is the Horvitz-Thompson (HT) for the total Y where $\mathbf{y} = [y_k] \in \mathbb{R}^{N\times 1}$. The HT estimator is a linear function of the random elements S_k of $\mathbf{S}$ for $k \in U$ since it can be expressed as

$$\hat{Y}_{HT} = \lambda_1 S_1 + ... + \lambda_N S_N, \tag{4.23}$$

where $\lambda_k = \frac{y_k}{\pi_k}$ for $k \in \{1,...,N\}$.[20] The error of $\mathrm{f}(\mathbf{S})$ is the variance of $\mathrm{f}(\mathbf{S})$ defined as

$$\mathbb{V}(\mathrm{f}(\mathbf{S})) = (\mathbf{d} \odot \mathbf{y})^{\mathrm{T}} \boldsymbol{\Delta} (\mathbf{d} \odot \mathbf{y}), \tag{4.24}$$

which is the quadratic function $Q: \mathbb{R}^N \to \mathbb{R}$, $Q_{\mathbf{A}}(\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z}$ with $\mathbf{A} = \boldsymbol{\Delta}$ and $\mathbf{z} = \mathbf{d} \odot \mathbf{y}$.

[20] Note the focus on linear functions of the random variables S_k for $k \in \{1,...,N\}$ instead of linear combination of the outcome y_k (Wolter, 2017).

Now, we add an adjustment through the scalar $\hat{\Gamma}_1 \in \mathbb{R}$ with the PA adjustment, $\hat{\Gamma}_1 = \frac{N}{\mathbf{d}^{\mathrm{T}}\mathbf{S}}$. We define the scalar-to-scalar valued function $\mathrm{f}^* : \mathbb{R} \to \mathbb{R}$ of $\mathbf{S}$ as

$$\mathrm{f}^*(\mathbf{S}) = \hat{\Gamma}_1 \mathbf{d}^{\mathrm{T}} (\mathbf{y} \odot \mathbf{S}) = N\left(\mathbf{d}^{\mathrm{T}}\mathbf{S}\right)^{-1} (\mathbf{d} \odot \mathbf{y})^{\mathrm{T}} \mathbf{S}. \tag{4.25}$$

Before proceeding, we verify that the sequence of estimators adjusted by a sequence of adjustment $\hat{\Gamma}_1$, $\mathrm{f}^*(\mathbf{S}_N)$, is consistent, or $\mathbb{E}\left(\mathrm{f}^*(\mathbf{S}_N) - \mathrm{f}(\mathbf{S}_N) -\right) = \mathcal{O}\left(n^{-1}\right)$ as $N \to \infty$.

The new function $\mathrm{f}^*(\mathbf{S})$ is the Hájek estimator (HJ) of the total Y, and since $\mathrm{f}^*(\mathbf{S})$ is nonlinear (of $\mathbf{S}$), the propagation error in $\mathrm{f}^*(\mathbf{S})$ is approximated using the first order approximation of the multivariate Taylor expansions of f^* evaluated at $\mathbf{S} = \boldsymbol{\pi}$ by

$$\mathbb{V}\left(\mathrm{f}^*(\mathbf{S})\right) \approx \left.\frac{\partial \mathrm{f}^*(\mathbf{S})^{\mathbf{T}}}{\partial \mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}} \boldsymbol{\Delta} \left.\frac{\partial \mathrm{f}^*(\mathbf{S})}{\partial \mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}}, \tag{4.26}$$

where $\frac{\partial \mathrm{f}^*(\mathbf{S})}{\partial \mathbf{S}} = \left(\frac{\partial \mathrm{f}^*(\mathbf{S})}{\partial S_1}, ..., \frac{\partial \mathrm{f}^*(\mathbf{S})}{\partial S_N}\right) \in \mathbb{R}^{1\times N}$ is the vector of the directional derivative[21] taken with respect to a vector field $\mathbf{S}$. Let $\mathcal{D} = \frac{\partial \mathrm{f}^*(\mathbf{S})}{\partial \mathbf{S}}$, then

[21] Vector derivatives are important in theoretical and applied physics as they arise in fields such as electricity, magnetism, and fluid mechanics among other areas. These are tools to study random fields in matrix representation.

$$\boldsymbol{\mathcal{D}} = \left[d_k y_k - \frac{Y}{N} d_k \right]_{k \in \{1,\ldots,N\}}. \tag{4.27}$$

As a vector of partial derivatives, each element of $\boldsymbol{\mathcal{D}}$, $\mathcal{D}_k$ for $k \in \{1,\ldots,N\}$, measures the change of $\mathrm{f}^*(\mathbf{S})$ with respect to S_k while S_l for $l \neq k \in \{1,\ldots,N\}$ remain constant. Since the function $\mathrm{f}^*(\mathbf{S})$ is very simple, the change is the same for all S_k. However, this observation establishes a link to replication methods for estimating variances such as the Jackknife, where each replicate measures the effect of the estimator when one element is removed. When considering the random variable $\mathbf{S}$, the Taylor variance resembles the replication methods because the variance is computed as a function of the changes in $\mathrm{f}^*(\mathbf{S})$ for each S_k, keeping the effect of the others constant. Each element in $\mathcal{D}_k$ can be viewed as a "replicate."

After algebraic simplification, the propagation error or variance of the function $\mathrm{f}^*(\mathbf{S})$ is

$$\mathbb{V}\left(\mathrm{f}^*(\mathbf{S})\right) \approx (\mathbf{d} \odot \mathbf{e})^T \boldsymbol{\Delta} (\mathbf{d} \odot \mathbf{e}) = Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{e}), \tag{4.28}$$

where $\mathbf{e} = \mathbf{y} - \bar{Y}$ is the vector of the residuals around the population mean $\bar{Y}$.

There is no easy way to compare the quadratic forms $Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{e})$ and $Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{y})$ for designs other than SRS to determine if the adjustment reduces the propagation of errors. However, we can determine an inequality that bounds the differences between

the two estimators. The upper bounds provide insights on the conditions when one estimator is more efficient than the other.

Let $\mathbf{Q}\Lambda\mathbf{Q}^{\mathrm{T}} = \mathbf{\Delta}$ be the spectral decomposition of $\mathbf{\Delta}$, where $\Lambda = \mathrm{diag}(\boldsymbol{\lambda})$ is the diagonal matrix of the vector of eigenvalues $\boldsymbol{\lambda} \in \mathbb{R}^N$, and $\mathbf{Q} = (\mathbf{q}_1, ..., \mathbf{q}_k)$ is the matrix of the eigenvectors $\mathbf{q}_k \in \mathbb{R}^{N \times 1}$ for $k \in U$. Since by definition $\mathbf{\Delta}$ is symmetric and positive semidefinite (e.g., it is a fixed size sample design, see Section 5.3), all eigenvalues except for one are real positive numbers, and the matrix $\mathbf{Q}$ is orthogonal with rows and columns forming an orthonormal basis. Then the quadratic forms (4.22) and (4.28) can be written as

$$Q_{\mathbf{\Delta}}(\mathbf{d} \odot \mathbf{y}) = \sum_{k \in U} \lambda_k \left\| \mathbf{q}_k^{\mathrm{T}} (\mathbf{d} \odot \mathbf{y}) \right\|^2 \text{ and}$$
$$Q_{\mathbf{\Delta}}(\mathbf{d} \odot \mathbf{e}) = \sum_{k \in U} \lambda_k \left\| \mathbf{q}_k^{\mathrm{T}} (\mathbf{d} \odot \mathbf{e}) \right\|^2, \tag{4.29}$$

which are weighted sums of the squared $L - 2$ norms (i.e., Euclidean norm) of the projections of $\mathbf{d} \odot \mathbf{y}$ or $\mathbf{d} \odot \mathbf{e}$ to the eigenvectors of the matrix $\mathbf{Q}$ where the weights are the eigenvalues λ_k for $k \in U$. Note that the only difference in the quadratic forms $Q_{\mathbf{\Delta}}(\mathbf{d} \odot \mathbf{y})$ and $Q_{\mathbf{\Delta}}(\mathbf{d} \odot \mathbf{e})$ is the variables $\mathbf{y}$ and $\mathbf{e}$ since both have the same set of eigenvalues, orthonormal basis, and sampling weights (e.g., the matrix $\mathbf{\Delta}$ is the same in both). The expressions can easily be evaluated for simple random designs. The comparison is not as straightforward in informative designs where there is an interaction of $\mathbf{y}$ or $\mathbf{e}$ and the sample design represented by $\mathbf{Q}$.

To have a general sense of the differences between the estimators, let $\lambda_{\max}$ be the maximum eigenvalue of $\boldsymbol{\Delta}$ defined as $\lambda_{\max} = \arg\max_{\lambda \in \Lambda} \{\lambda_1, ..., \lambda_N\}$, then the following inequalities hold

$$Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{y}) \leq \lambda_{\max} \|\mathbf{d}\|^2 \|\mathbf{y}\|^2 \text{ and}$$
$$Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{e}) \leq \lambda_{\max} \|\mathbf{d}\|^2 \|\mathbf{e}\|^2. \tag{4.30}$$

Since $\lambda_{\max}$ and $\|\mathbf{d}\|^2$ are always positive, the ratio of the quadratic forms is

$$\frac{Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{e})}{Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{y})} \leq \frac{\|\mathbf{e}\|^2}{\|\mathbf{y}\|^2}. \tag{4.31}$$

The ratio in (4.31) shows that the propagation error in $\mathrm{f}^*(\mathbf{S})$ is smaller than $\mathrm{f}(\mathbf{S})$ if the sum of the squared residuals or $\mathbf{e}^{\mathrm{T}}\mathbf{e}$ is smaller than the sum of squared y values or $\mathbf{y}^{\mathrm{T}}\mathbf{y}$. This expression is similar to the ratio of partitioned sums of squares. If we let y be related to π as $y_k = \beta\pi_k$, then the ratio of the variances is

$$\frac{Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{e})}{Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{y})} \leq \frac{\sum_{k \in U} \pi_k^2 + N - 2n}{\sum_{k \in U} \pi_k^2}. \tag{4.32}$$

The ratio in (4.32) shows that the HT estimator is more efficient than the HJ estimator when y is a linear function of π. The adjustment $\hat{\Gamma}_1$ increases the propagation error in $\mathrm{f}^*(\mathbf{S})$ compared to $\mathrm{f}(\mathbf{S})$. This situation (a high linear correlation) is common in practice. In Example 1.1 on page 7, the hospitals are drawn using the number of beds as the hospital measure of size, and the number of beds is correlated to the outcome variable for hospital expenditures (e.g., larger hospitals measured in terms of the

number of beds have large expenditures). In this case, the HT estimator for total expenditures is more efficient than the HJ estimator as shown in Table 1.5.

Using the same approach, we examine the propagation error for the ratio estimator (RA) compared to the HT estimator where

$$\hat{Y}_{RA} = X \frac{\mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{S})}{\mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{S})}. \tag{4.33}$$

Now, we add the PA adjustment through the scalar $\hat{\Gamma}_X = \frac{X}{\hat{X}_{HT}} = \frac{X}{\mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{S})} \in \mathbb{R}$ to the function in (4.22). We define the scalar-to-scalar valued function $\mathrm{f}^R : \mathbb{R} \to \mathbb{R}$ of $\mathbf{S}$ as

$$\mathrm{f}^R(\mathbf{S}) = \hat{\Gamma}_X \, \mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{S}) = X \frac{\mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{S})}{\mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{S})}. \tag{4.34}$$

Equation (4.34) is a nonlinear function of $\mathbf{S}$ so the variance is approximated by (4.26). After algebraic simplification, the propagation error of $\mathrm{f}^R(\mathbf{S})$ is

$$\mathbb{V}\left(\mathrm{f}^R(\mathbf{S})\right) \approx (\mathbf{d} \odot \mathbf{e})^T \, \boldsymbol{\Delta} (\mathbf{d} \odot \mathbf{e}) = Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{e}), \tag{4.35}$$

where $\mathbf{e} = \mathbf{y} - \mathbf{x}\frac{Y}{X}$ and $Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{e})$ is the quadratic form of the vector $\mathbf{d} \odot \mathbf{e}$ and the matrix $\boldsymbol{\Delta}$.

The vector $\boldsymbol{\mathcal{D}} = \frac{\partial \mathrm{f}^R(\mathbf{S})}{\partial \mathbf{S}}$ can be written as

$$\boldsymbol{\mathcal{D}} = \left[d_k y_k - d_k x_k \frac{Y}{X} \right]_{k \in \{1,\ldots,N\}}, \tag{4.36}$$

which shows the effect of changes if $\mathrm{f}^R(\mathbf{S})$ for each S_k. In this case, the change depends on the value of x_k. The variance is the sum of the cross-product of all these "replicates" $\mathcal{D}_k$ for $k \in \{1,\ldots,N\}$. The similarities between the Taylor series "replicates" and replication methods are also observed here.

The ratio of the quadratic forms using the upper bounds of $\boldsymbol{\Delta}$ is

$$\frac{Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{e})}{Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{y})} \leq \frac{\|\mathbf{e}\|^2}{\|\mathbf{y}\|^2}. \tag{4.37}$$

Assume that the outcome is a constant, $y_k = c$ for $k \in \{1,..,N\}$. After simplifying (4.37), the ratio of the quadratic forms is

$$\frac{Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{e})}{Q_{\boldsymbol{\Delta}}(\mathbf{d} \odot \mathbf{y})} \leq (1-N)^2. \tag{4.38}$$

This ratio is always greater than one, and the value is very large due to the assumption of constant outcomes which does not occur in practice. Although this assumption does not hold in practice, this result shows that when y is not correlated to x, the ratio estimator can be very inefficient compared to the HT estimator.

4.8 Incorporating Population Totals into the Pseudo-Likelihood

The second motivation for the PA estimators is to improve the precision of the PML estimates by incorporating the additional information represented by the control totals of the auxiliary variables directly in the PL function. Until now, the PML approach has been used mainly to estimate the model parameters instead of finite population characteristics, and the auxiliary variable population totals are not used in this approach (Binder, 1983; Binder & Roberts, 2009).

Incorporating the auxiliary population information is based on the following observations. Assume a linear superpopulation model $\mathcal{M}$, where $\mathcal{N}\left(\eta_\beta, \sigma^2\right)$ with $\eta_\beta = \mathbf{x}\boldsymbol{\beta}$. When this model is fitted to the finite population, the MLE of the regression coefficients $\boldsymbol{\beta}_{mle} \in \mathbb{R}^{1\times P}$ meet the following condition

$$\hat{\boldsymbol{\beta}}_{mle} = \mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{T}_{\mathbf{xy}}, \tag{4.39}$$

where $\mathbf{T}_{\mathbf{xx}} = \mathbf{x}^{\mathrm{T}}\mathbf{x} = \left[\sum_{k\in U} x_{ik}x_{jk}\right] \in \mathbb{R}^{P\times P}$ and $\mathbf{T}_{\mathbf{xy}} = \mathbf{x}^{\mathrm{T}}\mathbf{y} = \left[\sum_{k\in U} x_{ik}y_{jk}\right] \in \mathbb{R}^{P\times 1}$. Let the first component in $\mathbf{x}_k$ be one for $k \in \{1,..,N\}$. Let $\mathbf{r}_1 \in \mathbb{R}^{1\times P}$ be the first partitioned row, and $\mathbf{c}_1 \in \mathbb{R}^{P\times 1}$ be the first partitioned column of the matrix $\mathbf{T}_{\mathbf{x},\mathbf{x}}$. The elements of $\mathbf{r}_1$ and $\mathbf{c}_1$ correspond to a vector of the auxiliary variable population totals.

$$\mathbf{r}_1 = \mathbf{c}_1^{\mathrm{T}} = \mathbf{X} = \left(N, X_1, ..., X_{P-1}\right) \in \mathbb{R}^{1\times P}. \tag{4.40}$$

When the model is fitted to a sample drawn according to a sample design $p\left(A = a\right)$ (see Definition 1.5), the sample-based or PML estimator of $\hat{\boldsymbol{\beta}}_{mle}$ is

$$\hat{\boldsymbol{\beta}}_{mle} - \hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\,\hat{\mathbf{T}}_{\mathbf{xy}} \mid \mathcal{F} = \mathcal{O}_p\left(n^{-1}\right), \tag{4.41}$$

where $\hat{\mathbf{T}}_{\mathbf{xx}}$ is the sample-based estimator of $\mathbf{T}_{\mathbf{xx}}$ given by

$$\hat{\mathbf{T}}_{\mathbf{xx}} = \left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{x} = \left[\sum_{k\in U} d_k x_{pk} x_{qk} S_k\right]_{p,q\in\{1,\ldots,P\}} \in \mathbb{R}^{P\times P}. \tag{4.42}$$

$\hat{\mathbf{T}}_{\mathbf{xy}} = \left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{y} = \left[\sum_{k\in U} d_k x_{pk} y_k S_k\right]_{p\in\{1,\ldots,P\}} \in \mathbb{R}^{P\times 1}$ is the sample-based estimator of $\mathbf{T}_{\mathbf{xy}}$. The sample-based estimators of $\mathbf{r}_1$ and $\mathbf{c}_1$ of $\mathbf{T}_{\mathbf{xx}}$ are the first row and column of $\hat{\mathbf{T}}_{\mathbf{xx}}$ are given by

$$\hat{\mathbf{r}}_{HT,1} = \hat{\mathbf{c}}_{HT,1}^{\mathrm{T}} = \hat{\mathbf{X}}_{HT} = \left(\hat{N}_{HT}, \hat{X}_{HT,1}, ..., \hat{X}_{HT,P-1}\right) \in \mathbb{R}^{1\times P}, \tag{4.43}$$

where $\mathbf{r}_1 - \hat{\mathbf{r}}_{HT,1} = \mathcal{O}_p\left(n^{-1}\right)$ and $\mathbf{c}_1 - \hat{\mathbf{c}}_1 = \mathcal{O}_p\left(n^{-1}\right)$. However, the population totals $\mathbf{X}$ are known, and there is no need to use estimates in $\mathbf{r}_1$ and $\mathbf{c}_1$ of $\hat{\mathbf{T}}_{\mathbf{xx}}$. Excluding the population totals $\mathbf{X}$ from $\hat{\mathbf{T}}_{\mathbf{xx}}$ does not take advantage of all the information available.

There are different ways to incorporate the population totals $\mathbf{X}$ in $\hat{\mathbf{T}}_{\mathbf{xx}}$, which, at the same time, incorporates them into $\hat{\boldsymbol{\beta}}_{pmle}$. One method is through the PA adjustment factor which is a diagonal matrix $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}} \in \mathbb{R}^{P\times P}$ defined as

$$\hat{\boldsymbol{\Gamma}}_{\mathbf{X}} = \mathbf{D}_{\mathbf{X}}\mathbf{D}_{\hat{\mathbf{X}}}^{-1}. \tag{4.44}$$

Then the PA adjusted estimator of $\mathbf{T}_{\mathbf{xx}}$, $\hat{\mathbf{T}}_{\mathbf{xx},PA} \in \mathbb{R}^{P\times P}$, is

$$\hat{\mathbf{T}}_{\mathbf{xx},PA} = \hat{\mathbf{T}}_{\mathbf{xx}}\hat{\boldsymbol{\Gamma}}_{\mathbf{X}} = \hat{\mathbf{T}}_{\mathbf{xx}}\mathbf{D}_{\mathbf{X}}\mathbf{D}_{\hat{\mathbf{X}}}^{-1} = \mathbf{D}_{\mathbf{X}}\hat{\mathbf{T}}_{\mathbf{xx}}\mathbf{D}_{\hat{\mathbf{X}}}^{-1}. \tag{4.45}$$

The PA adjustment $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$ removes the sampling variability from $\hat{\mathbf{r}}_{1,PA}$, and $\hat{\mathbf{c}}_{1,PA}$ (e.g. $\mathrm{E}\left(\mathbf{r} - \hat{\mathbf{r}}_{1,PA} \mid \mathcal{F}\right) = \mathrm{E}\left(\mathbf{c}^{\mathrm{T}} - \hat{\mathbf{c}}_{1,PA}^{\mathrm{T}} \mid \mathcal{F}\right) = \mathbf{0} \in \mathbb{R}^{1\times P}$). The propagation of the adjustment $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$ also reduces the variability of other elements of $\hat{\mathbf{T}}_{\mathbf{xx},PA}$. To examine the propagation errors, we rewrite $\hat{\mathbf{T}}_{\mathbf{xx},PA}$ in terms of their elements as

$$\hat{\mathbf{T}}_{\mathbf{xx},PA} = \left[\hat{T}_{x_p x_q}\frac{X_p}{\hat{X}_p}\frac{X_q}{\hat{X}_p}\right]_{p,q\in\{1,\ldots,P\}} \in \mathbb{R}^{P\times P}. \tag{4.46}$$

Although the effect of $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$ is similar to calibrating to the population totals of the elements in $\hat{\mathbf{r}}_{1,PA}$, and $\hat{\mathbf{c}}_{1,PA}$, the remaining adjusted entries of $\hat{\mathbf{T}}_{\mathbf{x},\mathbf{x},PA}$ do not meet the calibration restriction since the population total of $\hat{T}_{x_p x_q}$ is not $X_p X_q$.

This type of PA adjustment for these entries is justified as a special class of improved estimators proposed by Srivastava & Jhajj (1981). They define this class of estimators adjusted by the product of two estimators:

$$\hat{Y} = \hat{Y}\,\mathrm{H}(u,v), \tag{4.47}$$

where $u = \dfrac{\hat{\mathbf{X}}_i}{\mathbf{X}_i}$, $v = \dfrac{\hat{\mathbf{X}}_j}{\mathbf{X}_j}$ and $y_k = x_{ki}x_{kj}$, $\mathrm{H}(u,v)$ is a function of u and v such that

1. The point (u,v) assumes the value in a closed convex subset in $\mathbb{R}^2$ containing the point $(1,1)$;
2. The function $\mathrm{H}(u,v)$ is continuous and bounded in $\mathbb{R}^2$;
3. $\mathrm{H}(1,1) = 1$; and
4. The first and second order partial derivatives of $\mathrm{H}(u,v)$ exist and are continuous.

The properties of this class of estimators, such as asymptotic bias and MSE, are described in Srivastava & Jhajj (1981).

The idea of adjusting for estimators using products of auxiliary variables is the motivation for creating alternative versions of PA estimators.

4.9 Alternative Forms of PA Estimators

Before describing the methods to incorporate population characteristics other than the population total of the PA estimator, we derive the PA estimator of the total of y with a superpopulation model $\mathcal{M}_y$ where $y_k \mid x_k \overset{iid}{\sim} \mathcal{N}\left(\dfrac{\beta}{x_k}, \dfrac{\sigma_0^2}{x_k}\right)$. We assume that

only the auxiliary population totals (N, X) are available. After solving the PL for the model $\mathcal{M}_y$ fitted to the observed sample to obtain the PMLE of β, we obtain

$$\hat{\beta}_{pmle} = \frac{\sum_{k \in A} d_k y_k}{\sum_{k \in A} \frac{d_k}{x_k}}. \tag{4.48}$$

The auxiliary variable is $\frac{1}{x_k}$, so the population total is $\sum_{k \in U} \frac{1}{x_k}$. Since we assume that we do not have the entire population, the population total for this variable cannot be computed. As an alternative, we propose a PA adjustment for the total $\frac{1}{X}$ with the sample-based estimate defined as $\frac{1}{\hat{X}_{HT}}$. Then the PA adjustment for this estimator is

$$\Gamma_{1/X} = \frac{1/X}{1/\hat{X}_{HT}} = \frac{\hat{X}_{HT}}{X}. \tag{4.49}$$

Note that $\frac{1}{X} - \frac{1}{\hat{X}_{HT}} = \mathcal{O}_p\left(\frac{1}{n}\right)$.

The PA estimator is then obtained applying the adjustment to (4.49) and plugging into the generic PA estimator (1.25). The PA estimator for the total Y for this model is

$$\hat{Y}_{PA} = \hat{Y}_{HT} \frac{\hat{X}_{HT}}{X}. \tag{4.50}$$

The estimator in (4.50) is the generalization of the *product ratio estimator* proposed by Murthy (1964). Although the product ratio estimator is a PA estimator with population totals that do not quite match the auxiliary variables, the important point is

that estimators can be derived using any adjustments as long are they are correlated to the outcome. This observation provides some alternatives for PA estimators.

The PA estimator described in the previous chapters has the form of a ratio estimator based on ratios to totals as

$$\hat{\Gamma}_{X_p} = \frac{X_p}{\hat{X}_{HT,p}}, \tag{4.51}$$

for $p \in \{1,\ldots,P\}$. This estimator is called the *total ratio PA estimato*r. The alternative is based on the inverse of $\hat{\Gamma}_{x_p}$, and applies to product ratio estimators described above.

An alternative is a PA adjustment based on the ratio of the population means to the sample-based estimate of the same mean; this is called the *mean ratio PA estimator*. The PA adjustment for the mean ratio is

$$\hat{\Gamma}_{X_p} = \frac{\bar{X}_p}{\bar{\hat{X}}_{HJ,p}}, \tag{4.52}$$

where $\bar{X}_p = \frac{X_p}{N}$ and $\bar{\hat{X}}_{HJ,p} = \frac{\hat{X}_{HT,p}}{\hat{N}_{HT}}$ for $p \in \{1,\ldots,P\}$.

For sample designs where $\hat{N}_{HT} = N$, the total ratio PA estimator and mean ratio PA estimator produce the same estimator. Otherwise, there are differences in the estimators due to the different adjustments made to the regression coefficients. For example, if x_1 is the term for the intercept, the PA adjustment for this term is always

one for the mean ratio PA estimator, but the adjustment affects the slope regression coefficients. In contrast, using the total ratio estimator, the adjustment is $\frac{N}{\hat{N}_{HT}}$, and the variation $\hat{N}_{HT}$ affects the coefficients of the slopes.

The third group does not rely on the population totals represented as the sum of the elements in the frame. Instead, the estimators in this group use an function of the expected value as the factor. For example, the total ratio estimator is the exponential mean ratio PA estimator, with PA adjustment factor defined as

$$\hat{\Gamma}_{X_p} = \exp\left(\frac{\bar{X}_p}{\bar{X}_{HJ,p}}\right), \tag{4.53}$$

for $p \in \{1,\ldots,P\}$. An exponential total ratio could also be computed by replacing the means by totals. There are the corresponding alternatives for product estimators.

If the population variance is available, a PA estimator can be computed as

$$\hat{\Gamma}_{X_p S^2_p} = \frac{\bar{X}_p}{\bar{X}_{HJ,p}} \frac{S_p^2}{S_{HJp}^2}, \tag{4.54}$$

where $S_p^2 = \frac{\sum\left(x_k - \bar{X}\right)}{N-1}$ is the population variance of X_P and $\hat{S}_{HJ,p}^2 = \frac{\sum d_k \left(x_k - \hat{\bar{X}}_{HJ,p}\right)}{\hat{N}_{HT} - 1}$ is the sample-based estimate of the population variance S_p^2 for $p \in \{1,\ldots,P\}$.

Many other estimators can be constructed in this way based on the product of the population coefficient variation, population kurtosis, and population median. Estimators that are ratios to other population characteristics, such as

$$\hat{\Gamma}_{\bar{X}_p+C_{x_p}} = \frac{\bar{X}_p + C_{x_p}}{\hat{\bar{X}}_p + \hat{C}_{x_p}}, \tag{4.55}$$

where $C_{x_p} = \frac{S_{x_p}}{\bar{X}_p}$ is the population coefficient of variation of X_p for $p \in \{1,\ldots,P\}$ could also be constructed. The difficulty lies in the fact that it is unusual to know these population quantities.

The PA adjustment using population characteristics described above is similar to a regression coefficient that is constant for all the cases in the sample. We consider the same population characteristics but use the information at the sample level. The population characteristics that can be incorporated at the sampled element level are listed in Table 4.6. The table shows the auxiliary variable and the population totals for these population characteristics.

EXAMPLE 4.3. Let y be the variable of interest with a superpopulation model $\mathcal{M}_y$ where $y_k \overset{iid}{\sim} N\left(\beta_0 + \beta_x x_k + \beta_z z_k, \sigma^2\right)$, $z_k = \frac{\left(x_k - \bar{X}\right)^2}{N}$, and the population totals $\left(N, X, Z\right)$ where $Z = S_X^2$. The linear PA estimator for the total Y for this model is

$$\hat{Y}_{PA} = N\hat{\beta}_{pmle,0} + X\hat{\beta}_{pmle,X} + S_X^2\hat{\beta}_{pmle,Z}\,. \tag{4.56}$$

Note that the estimate of the total of the auxiliary variable z_k is $\widehat{S_X^2} = \sum_{k\in A} d_k \hat{z}_k = \sum_{k\in A} \frac{\left(x_k - \hat{\bar{X}}_{HJ}\right)^2}{\hat{N}_{HT} - 1}$ and the population total is $S_X^2 = \sum_{k\in U} \frac{\left(x_k - \hat{\bar{X}}\right)^2}{N-1}$.

Table 4.6 Auxiliary variables and population totals for population characteristics at the sampled element level

Population characteristic	Working model	Auxiliary variable z_k	Population total Z	PA adjustment factor $\hat{\Gamma}_X = \frac{Z}{\hat{Z}}$
Variance	$y_k \overset{iid}{\sim} \mathcal{N}\left(z_k\beta, \sigma^2\right)$	$\frac{\left(x_k - \hat{\bar{X}}_{HJ}\right)^2}{\hat{N}_{HT} - 1}$	S_X^2	$\frac{S_X^2}{\widehat{S_X^2}}$
Quantile*	$y_k \overset{iid}{\sim} \mathcal{N}\left(z_k\beta, \sigma^2\right)$	$\delta\left(x_k \le \hat{Q}_x\left(P_0\right)\right) / \hat{N}_{HT}$	$Q_X\left(P_0\right) = X_0$	$\frac{Q_X\left(P_0\right)}{\widehat{Q_X\left(P_0\right)}}$
Coefficient of variation	$y_k \overset{iid}{\sim} \mathcal{N}\left(z_k\beta, \sigma^2\right)$	$\frac{\left(x_k - \hat{\bar{X}}_{HJ}\right)^2}{\left(\hat{N}_{HT} - 1\right)\hat{\bar{X}}_{HT}^2}$	CV_X^2	$\frac{CV_X^2}{\widehat{CV_X^2}}$
Kurtosis	$y_k \overset{iid}{\sim} \mathcal{N}\left(z_k\beta, \sigma^2\right)$	$\frac{\left(x_k - \hat{\bar{X}}_{HJ}\right)^4}{\left(\hat{N}_{HT} - 1\right)\left(\widehat{S_{HT}^2}\right)^2}$	K_X	$\frac{K_X}{\widehat{K_X}}$
Skewness	$y_k \overset{iid}{\sim} \mathcal{N}\left(z_k\beta, \sigma^2\right)$	$\frac{\left(x_k - \hat{\bar{X}}_{HJ}\right)^3}{\left(\hat{N}_{HT} - 1\right)\left(\widehat{S_{HT}^2}\right)^{3/2}}$	G_X	$\frac{G_X}{\widehat{G_X}}$

*Note: $\delta\left(x_k \le \hat{Q}_x\left(P_0\right)\right) = 1$ if $x_k \le \hat{Q}_x\left(P_0\right)$, 0 otherwise, $\hat{Q}_x\left(P_0\right) = \hat{F}^{-1}\left(P_0\right)$ where $\hat{F}\left(x_{P_o}\right) = \frac{\sum_{k \in A} d_k 1_{\left\{x \le x_{P_o}\right\}}}{\hat{N}_{HT}}$.

REMARK 4.1 As in the PA estimator with the PA adjustment factor

$$\hat{\Gamma}_{X_p} = \frac{X_p}{\hat{X}_{HT,p}},$$

the role of the adjustment factors for alternative PA estimators in (4.52) (4.53), (4.54) and those listed in the last column Table 4.6 is to incorporate the auxiliary variable population information (e.g., population mean, total, coefficient of variation, variance) into the PL. As in the PA estimator, these adjustments are expected to reduce the variance of the estimator if the auxiliary variables are related to the outcome variable.

Chapter 5 Deriving the Asymptotic Properties of Survey Sampling Estimators

In this chapter, we derive the asymptotic properties of the parametric (PA) estimator. Most estimators proposed in the survey sampling literature derive their large sample properties by establishing an asymptotic equivalence to the Horvitz-Thompson (HT) estimator (see, for example, Wu & Sitter, 2001; Breidt & Opsomer, 2017). If the proposed estimator is asymptotically equivalent to the HT estimator, then it inherits the HT asymptotic properties. The HT estimator is design consistent, and the sequence of estimators $Z_N = \frac{\hat{\bar{Y}}_{HT,N} - \bar{Y}_N}{\sqrt{\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{HT,N}\right)}}$ converges in distribution to $\mathcal{N}(0,1)$ in a sequence of increasing size finite populations (N) and samples sizes (n). Thus, the proposed estimator is also consistent with a limiting normal distribution. Using similar relationships, the asymptotic design variance of the proposed estimator is equivalent to the asymptotic design-based variance of the HT estimator of the residuals $e_k = y_k - \hat{\mu}_k$ where $\hat{\mu}_k$ is the fitted mean of the model. This approach is not generally used in the classical asymptotic statistical literature for studying estimators defined as functions of random variables (Lehmann, 1999).

Although this approach is valid, it is not informative of the rate of convergence of the proposed estimator. For example, the proposed estimator might require large samples to approach its limit, and its performance may be very poor for small sample sizes. The current large sample approach used in survey sampling does not provide insights

into the proposed estimator's efficiency. Consequently, most papers include simulation studies to examine their properties empirically.

We take a different approach for the study of the estimator's large sample properties in the PA framework. One significant difference is the notation and algebra. We rely heavily on discrete multivariate statistics matrix notation, matrix operations, and matrix calculus (e.g., quadratic forms of matrices, matrix inequalities, eigenvalues, or vector-induced matrix norms). The main advantage is the ease of deriving the estimator's asymptotic properties.

The second difference is the focus on the random variables S_k, elements of the discrete random vector $\mathbf{S} = (S_1, ..., S_k, ..., S_N)$, with the sample membership indicators (see Definition 1.5 on page 44). This vector is the only stochastic component involved in the theory. This idea is an extension of the method proposed by Cornfield (1944) that enables the use of results from standard asymptotic theory to derive the statistical properties of finite population estimators. Further extending this idea to random vectors and matrices reduces the derivation of the formulas for expected values and variances, so it becomes a simple algebraic routine while providing new insights into the properties of the estimators.

We begin with the idea discussed by Tillé (2006), where any sample design can be uniquely described by the vector of the expected values, $\mathbb{E}(\mathbf{S}) = \boldsymbol{\pi}$ and the variance-covariance matrix of $\mathbf{S}$, $\mathbb{C}(\mathbf{S}) = \boldsymbol{\Delta}$. We show that the variance-covariance matrix $\boldsymbol{\Delta}$ has unique mathematical properties determined by sample design. Estimators such as

the Horvitz-Thompson (HT), Hájek (HJ), generalized regression (GREG), and parametric (PA) are defined as functions of the membership indicators of $\mathbf{S}$. The estimators or functions can be linear or nonlinear, and their asymptotic properties are systematically derived applying theorems of linear and nonlinear functions of sequences of random variables.

In the following sections, we discuss the foundations of different approaches to estimation from survey data with full response and show how any sample design is uniquely defined by a multivariate probability mass function of the discrete random vector $\mathbf{S}$ that defines the type of sample design. The matrix approach to the large sample properties of the estimator is then illustrated. This approach allows us to derive the expression of the estimator, its variance, and variance estimator, and their asymptotic properties.

5.1 Estimation Frameworks

Different theories for survey estimation depend on two random processes used to model the sample selection: one process is unobservable and generates the finite population from a superpopulation model; and the other is observable that selects the sample from the finite population. This setting is similar to the Rubin-Bleuer & Schiopu Kratina's probability product-space for the framework for joint design based and model-based inference. (Rubin-Bleuer & Schiopu Kratina, 2005).

The process that generates the finite population and draws the sample for the realized population is hierarchical. At the first stage, the finite population $\mathcal{F}$ with an outcome

variable y is generated as N identically independent distributed realizations (*iid*), $y_k \in U$, from a superpopulation model $\mathcal{M}_y$ with a distribution f_Y. In the second stage, a sample of size n is selected from the realized finite population, according to a sample design $p(\mathbf{S}=\mathbf{s})$ defined by a random vector $\mathbf{S}$ with a multivariate probability mass function $\mathrm{f}_{\mathbf{S}}$. Both variables are well defined with

$$\begin{aligned} \mathbf{y}_N \mid \mathbf{x}_N &\overset{iid}{\sim} \mathrm{f}_Y(\boldsymbol{\theta}), \text{ and} \\ \mathbf{S}_N \mid \mathbf{y}_N &\sim \mathrm{f}_{\mathbf{S}}(\boldsymbol{\pi}, \boldsymbol{\Delta}). \end{aligned} \tag{5.1}$$

Different estimation frameworks are the result of assumptions of the sampling distributions of $\mathbf{y}$ and $\mathbf{S}$. The estimation frameworks based on the random vectors $\mathbf{y}$ and $\mathbf{S}$ are listed in Table 5.1.

Table 5.1 Estimation frameworks as a function of random vectors $\mathbf{y}$ and $\mathbf{S}$

Estimation framework	Distribution	Source of variation	Target of Estimation	Comment
Design-based	$\mathrm{f}_{\mathbf{S}}(\mathbf{S} \mid \mathbf{Y}=\mathbf{y})$	$\mathbf{S}$, observed	Y	The variable $\mathbf{y}$ is fixed and considered as constant
Model-based	$\mathrm{f}_{\mathbf{Y}}\left(\mathbf{y} \mid \mathbf{S}=\mathbf{1}_{\{S_k=1\}}\right)$	$\mathbf{y}$, unobserved	Y	Sampling distribution of $\mathbf{S}$ is ignored
Super-population	$\mathrm{f}_{\mathbf{Y},\mathbf{S}}(\mathbf{y},\mathbf{S})$	$\mathbf{S}$, observed and $\mathbf{y}$ unobserved	θ	Both $\mathbf{y}$ and $\mathbf{S}$ are random variables

The differences among the estimation frameworks depend on how $\mathbf{y}$ and $\mathbf{S}$ are treated when producing estimates and inferences. Once this treatment is defined, it becomes straightforward to derive the statistical properties of the estimators in any of these frameworks.

REMARK 5.1. In the design-based approach, the random vector $\mathbf{S}$ is the only source of variability; all design-based estimators are functions of $\mathbf{S}$. In contrast, in the superpopulation approach, both $\mathbf{S}$ and $\mathbf{y}$ are random and contribute to the variability of the estimators, and the target of the estimator is not a finite population characteristic but a parameter θ of the superpopulation model. That is, there are two components of the variance, one from the finite population generation and the second from sample selection. For model-based estimation, the sample selection is ignored in estimation if the sample is balanced. Since in all frameworks the estimators are functions of these vectors of random variables, standard multivariate statistical tools can be used to derive their large sample properties. In the following sections, we focus only on the asymptotic properties of design-based estimators, that is, we condition on $\mathbf{y} = \mathbf{y}_0$ which becomes a vector of constants.

5.2 The Probability Mass Function of the Random Vector S

Sample designs $p(A = a)$ where A is some random subset of a population and a is a particular sample that was selected, can be uniquely defined as follows: let $\mathbf{S} \in \{0,1\}^N$ a vector-valued random variable with a discrete multivariate distribution

consisting of N random sample membership indicators $\mathbf{S} = (S_1,...,S_N)^{\mathrm{T}}$, with an expected value $\mathbb{E}(\mathbf{S} \mid \mathcal{F}) = \boldsymbol{\pi}$ where $\boldsymbol{\pi} = [\pi_k] \in (0,1)^N$ is the vector of the first-order inclusion probabilities $\pi_k > 0$ [22] for $k \in U$, $\mathbb{C}(\mathbf{S} \mid \mathcal{F}) = \mathbb{E}\left(\mathbf{S}\mathbf{S}^{\mathrm{T}} \mid \mathcal{F}\right) - \boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}} = \boldsymbol{\Delta}$ is the variance-covariance matrix of $\mathbf{S}$, where $\boldsymbol{\Delta} = [\Delta_{kl}] = [\pi_{kl} - \pi_k \pi_l]$ for $k,l \in U$, and π_{kl} is the second order probability of inclusion of elements k and l. The covariance matrix $\boldsymbol{\Delta}$ is a Hermitian matrix (Dol, Steerneman, & Wansbeek, 1996), which implies it has specific properties. $\boldsymbol{\Delta}$ is

(a.) A real (square) symmetric matrix;

(b.) A normal matrix such that $\boldsymbol{\Delta}\boldsymbol{\Delta}^{\mathrm{T}} = \boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Delta}$;

(c.) A matrix that can be diagonalized by a unitary matrix with real elements on the diagonal (finite-dimensional spectral theorem); and

(d.) A matrix with real and linearly independent eigenvalues.

Additional properties of $\boldsymbol{\Delta}$ depend on the type of sample design.

5.3 Types of Sample Designs

We are interested in discrete random vectors $\mathbf{S}$ such that $\mathbb{E}(\mathbf{S} \mid \mathcal{F}) = \boldsymbol{\pi} \in (0,1)^N$ and $\mathbb{C}(\mathbf{S} \mid \mathcal{F}) = \boldsymbol{\Delta}$. We also require $\pi_k > 0$ for all π_k in $\boldsymbol{\pi}$, and $\pi_{kl} > 0$ in

[22] In order to be a Lebesgue measure, $\pi_k > 0$.

$\boldsymbol{\Pi} = \left[\pi_{kl}\right] \in \mathbb{R}^{N \times N}$, where $\boldsymbol{\Pi}$ is the matrix with the second order of probability of inclusion, π_{kl}, for the elements k and l defined as the probability that the 2-tuple (k,l) are both selected in the sample. These conditions define a measurable design within the survey sampling theory context (Särndal, Swensson, & Wretman, 1992).

We use the variance of the sum of the elements of $\mathbf{S}$ to classify the sample designs. Let $Z : \mathbb{R}^N \mapsto \mathbb{R}$ be the function $Z = Z(\mathbf{S}) = \mathbf{1}^{\mathrm{T}}\mathbf{S}$, then Z represents the sum of all elements of $\mathbf{S}$. The variance of Z is $\mathbb{V}(Z \mid \mathcal{F}) = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Delta}\mathbf{1}$, and it can be decomposed as the sum of the contribution of the variances and covariance of the terms in $\mathbf{S}$ as

$$\mathbb{V}(Z \mid \mathcal{F}) = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Delta}\mathbf{1} = \sum_{k \in U} \mathbb{V}(S_k \mid \mathcal{F}) + \sum_{k,l \in U, k \neq l} \mathbb{C}(S_k, S_l \mid \mathcal{F}). \tag{5.2}$$

This expression has an intuitive meaning. Each element of $\mathbf{S}$, S_k, contributes to the total variance through the variance component, $\mathbb{V}(S_k \mid \mathcal{F})$, and through the sum of the covariances with the other elements $\sum_{l \in U, k \neq l} \mathbb{C}(S_k, S_l \mid \mathcal{F})$.

The value of $\mathbb{V}\left(\mathbf{1}^{\mathrm{T}}\mathbf{S} \mid \mathcal{F}\right)$ determines if it is a fixed sample size design or a random sample size design. This classification facilitates the derivation of the asymptotic properties of the estimators since these designs have very different properties of the variance-covariance matrix $\boldsymbol{\Delta}$.

5.3.1 Fixed Sample Size Designs

The random vector $\mathbf{S}$ represents a fixed sample size design if $\mathbb{V}\left(\mathbf{1}^{\mathrm{T}}\mathbf{S} \mid \mathcal{F}\right) = 0$. Some examples of fixed sample size designs are SRS, Sampford, Midzuno, and Tillé sampling (Tillé, 2006). These designs have the following properties:

(a) $\boldsymbol{\Delta}$ is positive semidefinite.

(b) If $\lambda_{\min}\left(\boldsymbol{\Delta}\right) \leq \lambda_{N-1} \leq \ldots \leq \lambda_2 \leq \lambda_{\max}\left(\boldsymbol{\Delta}\right)$ are the ordered eigenvalues of $\boldsymbol{\Delta}$, then $\lambda_{\min}\left(\boldsymbol{\Delta}\right) = 0$; that is, the eigenvalues $\lambda_k\left(\boldsymbol{\Delta}\right)$ for $k \in U$ are nonnegative.

(c) $\mathbf{1}\mathrm{row}_k\,\boldsymbol{\Delta} = 0$ and $\mathbf{1}^{\mathrm{T}}\mathrm{col}_k\,\boldsymbol{\Delta} = 0$ for $k \in U$, and $\mathrm{Tr}\left(\mathbf{I}\boldsymbol{\Delta}\right) = 0$, that is the sums of rows, the sum of columns, and the total sum of the elements of $\boldsymbol{\Delta}$ is zero.

(d) The sample size is computed as $n = \mathbf{1}^{\mathrm{T}}\boldsymbol{\pi}$.

5.3.2 Random Sample Size Designs

The discrete random vector $\mathbf{S}$ with parameters $\mathbb{E}\left(\mathbf{S} \mid \mathcal{F}\right) = \boldsymbol{\pi}$ and $\mathbb{C}\left(\mathbf{S} \mid \mathcal{F}\right) = \boldsymbol{\Delta}$ is a random sample size design if $\mathbb{V}\left(\mathbf{1}^{\mathrm{T}}\mathbf{S} \mid \mathcal{F}\right) \neq 0$. Some examples of random size designs are the Bernoulli, and PO (Tillé, 2006). Although this type of sampling is less frequently implemented in practice, random size designs are especially useful for modeling nonresponse. The additional properties of the random sample size designs are:

(a) $\boldsymbol{\Delta}$ is positive definite with all eigenvalues $\lambda_k\left(\boldsymbol{\Delta}\right) > 0$ for $k \in U$.

(b) $\boldsymbol{\Delta} = \text{diag}(\boldsymbol{\pi})$ because $\pi_{kl} = \pi_k \pi_l$ in $\boldsymbol{\Delta}$ for $k, l \in U : k \neq l$.

(c) The row and column sums are $\mathbf{1}^{\mathrm{T}} \text{row}_k \boldsymbol{\Delta} = \pi_k$, $\mathbf{1} \text{col}_l \boldsymbol{\Delta} = \pi_l$ for $k, l \in U$, and $\text{Tr}(\mathbf{I}\boldsymbol{\Delta}) = n$ where n is the expected sample size, $n = \mathbb{E}\left(\mathbf{1}^{\mathrm{T}} \mathbf{S} \mid \mathcal{F}\right)$.

(d) $\mathbb{V}\left(\mathbf{1}^{\mathrm{T}} \mathbf{S} \mid \mathcal{F}\right) = \mathbf{1}^{\mathrm{T}} \boldsymbol{\Delta} \mathbf{1} = \mathbf{1}^{\mathrm{T}} \left(\boldsymbol{\pi} \odot (1 - \boldsymbol{\pi})\right)$.

(e) Let $\mathbf{s} \in \{0,1\}^{N \times 1}$ be the vector of the realization of $\mathbf{S}$, $\mathbf{S} = \mathbf{s}$ then the observed sample size n_o is $n_o = \mathbf{1}^{\mathrm{T}} \mathbf{s}$.

(f) If $\lambda_{\min}(\boldsymbol{\Delta}) \leq \lambda_{N-1} \leq \ldots \leq \lambda_2 \leq \lambda_{\max}(\boldsymbol{\Delta})$ are the ordered eigenvalues of the variance-covariance matrix $\boldsymbol{\Delta}$, then the eigenvalues are the first order probability of inclusion $\boldsymbol{\pi}$. The largest eigenvalue of $\boldsymbol{\Delta}$, is $\lambda_{\max}(\boldsymbol{\Delta}) = \arg\max_{k \in U} \{\pi_k\}$.

REMARK 5.2 The properties and classification of sample designs based on the properties of variance-covariance matrix $\boldsymbol{\Delta}$ as a Hermitian matrix described above and in Sections 5.3, 5.3.1, and 5.3.2 have not been reported on the literature before.

5.4 Functions of the Random Vector S

We explore two basic functions of the random vector $\mathbf{S}$ using results from multivariate standard statistical limit theory to understand the statistical properties of design-based estimators.

5.5 Function for the Mean Vector of the Random Vectors S

Let $\mathbf{Z}:\mathbb{R}^N \to \mathbb{R}^N$ be a vector-valued function defined as $\mathbf{Z}(\mathbf{S}) = \frac{1}{N}\sum_{k=1}^{N}\mathbf{S}_k$ where $\mathbf{S}_k$ is the k-th realization of $\mathbf{S}$ for $k \in \{1,...,N\}$. The random vector $\mathbf{Z}$ is the average of all vectors $\mathbf{S}_k$. This function is a typical example found in statistical limit theory textbooks (e.g., Polansky, 2011). Define $\{\mathbf{Z}_N\}_{N=1}^{\infty}$ as the sequence of estimators $\mathbf{Z}$. Then

(a) $\lim_{N\to\infty}\mathbb{E}(\mathbf{Z}_N \mid \mathcal{F}) = \boldsymbol{\pi}$.

(b) $\mathbb{V}(\mathbf{Z}_N \mid \mathcal{F})$ is bounded, $\mathbb{V}(\mathbf{Z}_N \mid \mathcal{F}) = \mathcal{O}\left(\frac{1}{N}\right)$.

(c) Following from (a) and (b) $\{\mathbf{Z}_N\}_{N=1}^{\infty}$ is a consistent sequence of estimators of $\boldsymbol{\pi}$ (weak convergence, Polansky, 2011).

5.6 Function for the Mean of the Elements of the Random Vector S

Define the second function as follows: let $\mathrm{Z}:\mathbb{R}^N \to \mathbb{R}$ be a vector-to-scalar valued function $\mathrm{Z}(\mathbf{S}) = \frac{1}{N}\mathbf{1}^{\mathrm{T}}\mathbf{S}$. This function differs from the one in the previous section because Z is now the average of the N elements S_k of $\mathbf{S}$. The function Z is the overall sampling rate (or expected sampling rate in random sample size designs). To study the asymptotic properties of Z, let $\{Z_N\}_{N=1}^{\infty}$ be the sequence of estimators Z. The expected value and variance of this sequence are

$$\mathbb{E}(Z_N \mid \mathcal{F}) = \frac{1}{N}\mathbf{1}_N^{\mathrm{T}}\boldsymbol{\pi}_N, \text{ and} \tag{5.3}$$

$$\mathbb{V}(Z_N \mid \mathcal{F}) = \frac{1}{N^2}\mathbf{1}_N^{\mathrm{T}}\boldsymbol{\Delta}_N\mathbf{1}_N. \tag{5.4}$$

This function is not as common because the elements $S_k \in \mathbf{S}$ may not have the same expected value, $\mathbb{E}(S_k \mid \mathcal{F}) \neq \mathbb{E}(S_l \mid \mathcal{F})$ for $k \neq l$ and $k,l \in U$, and the 2-tuples (k,l) may be correlated (they are not independent).

Modified versions of asymptotic properties theorems for sequences of random variables that are neither identical nor independent are used to determine the asymptotic properties of this sequence. Furthermore, additional conditions on the behavior of the other parameters need to be imposed before deriving the asymptotic properties of the sequence of estimators $\{Z_N\}_{N=1}^{\infty}$. We discuss these conditions in more detail in Section 5.10.

The expressions (5.3) and (5.4) can be further simplified depending on the type of sample design. If $\mathbf{S}$ is a fixed sample size design, then $\mathbb{E}(Z \mid \mathcal{F}) = \frac{n}{N} = f$, where f is the overall sampling rate and $\mathbb{V}(Z \mid \mathcal{F}) = 0$. In this case, there is no need to find an upper bound for the sequence of estimators $\{Z_N\}_{N=1}^{\infty}$ because $\mathbb{V}(Z_N \mid \mathcal{F})$ is always zero.

In contrast, if $\mathbf{S}$ is a random sample size design, then the sequence $\{Z_N\}_{N=1}^{\infty}$ converges to the expected sampling rate $\frac{1}{N}\sum_{k \in N} \pi_k$. An upper bound of the variance $\mathbb{V}(Z_N \mid \mathcal{F})$ is found by applying regular rules for variances of random vectors, inequalities for quadratic forms of Hermitian matrices, and inequalities for eigenvalues in terms of matrix norms. So

$$\mathbb{V}(Z_N \mid \mathcal{F}_N) = \frac{1}{N^2}\mathbf{1}_N^{\mathrm{T}}\boldsymbol{\Delta}_N\mathbf{1}_N = \frac{1}{N^2}Q_{\boldsymbol{\Delta}_N}(\mathbf{1}_N) \leq \frac{1}{N^2}\lambda_{\max}(\boldsymbol{\Delta}_N)\|\mathbf{1}_N\|_2^2 = \frac{\lambda_{\max}(\boldsymbol{\Delta}_N)}{N}, \quad (5.5)$$

where $Q_{\boldsymbol{\Delta}_N}(\mathbf{1}_N) = \mathbf{1}_N^{\mathrm{T}}\boldsymbol{\Delta}_N\mathbf{1}_N$ is the quadratic form of the vector $\mathbf{1}_N$, $\lambda_{\max}(\boldsymbol{\Delta}_N)$ is the maximum eigenvalue of the matrix $\boldsymbol{\Delta}_N$, and $\|\mathbf{1}_N\|_2^2$ is the squared L^2-norm of the vector $\mathbf{1}_N$, where $\|\mathbf{1}_N\|_2^2 = \sum_{k \in N} 1^2 = N$. The variance $\mathbb{V}(Z_N \mid \mathcal{F}_N)$ is bounded by a function that depends on the largest eigenvalue of $\boldsymbol{\Delta}_N$, $\lambda_{\max}(\boldsymbol{\Delta}_N)$. In sample designs where the sample draws are independent (e.g., for $k \neq l, k, l \in U$), then

$\boldsymbol{\Delta}_N = \text{diag}\left(\boldsymbol{\pi}_N \odot \left(\mathbf{1}_N - \boldsymbol{\pi}_N\right)\right)$. Since for diagonal matrices, the eigenvalues are the elements of the diagonal, the largest eigenvalue is

$$\lambda_{\max}\left(\boldsymbol{\Delta}_N\right) = \max_{k \in U_N} \arg\left\{\Delta_{N,kk}\right\} = \max_{k \in U_N} \arg\left\{\pi_{N,k}\left(1-\pi_{N,k}\right)\right\}. \tag{5.6}$$

The bound of $\lambda_{\max}\left(\boldsymbol{\Delta}_N\right)$ depends on $\pi_{N,k}$. It is desirable to have a bound that does not depend on the first order inclusion probabilities. This bound can be found by noticing that $\lambda_{\max}\left(\boldsymbol{\Delta}_N\right)$ is the variance of a random variable with a Bernoulli distribution, which has a maximum value when $\pi = \frac{1}{2}$. Then, the variance of sequence $\left\{Z_N\right\}_{N=1}^{\infty}$ for designs with random sample sizes is bounded by

$$\mathbb{V}\left(Z_N \mid \mathcal{F}_N\right) \leq \frac{K_N}{N} = \mathcal{O}\left(\frac{1}{N}\right), \tag{5.7}$$

where $K_N = 0.5$. An implicit assumption in (5.7) is that $K_N = \mathcal{O}(1)$ which is true if $\lim_{N\to\infty} \lambda_{\max}\left(\boldsymbol{\Delta}_N\right) < \infty$.

5.7 Linear Functions of the Elements of the Random Vector S

We now introduce a constant vector $\mathbf{a} \in \mathbb{R}^N$ in the function Z. Let $\mathbf{a} = \left[a_k\right] \in \mathbb{R}^N$ be a vector of constants, and let $Z : \mathbb{R}^N \to \mathbb{R}$ be the function of $\mathbf{S}$ defined as $Z\left(\mathbf{S}\right) = \frac{1}{N}\mathbf{a}^{\mathrm{T}}\mathbf{S} = \frac{1}{N}\sum_{k=1}^{N} a_k S_k$. To study the asymptotic properties of this estimator, we

define the sequence of estimators $\{Z_N\}_{N=1}^{\infty}$ and apply the same rules as in Section **1.5**. The expected value and variance of $\{Z_N\}_{N=1}^{\infty}$ are

$$\mathbb{E}(Z_N \mid \mathcal{F}_N) = \frac{1}{N}\mathbf{a}_N^{\mathrm{T}}\boldsymbol{\pi}_N, \text{ and} \tag{5.8}$$

$$\mathbb{V}(Z_N \mid \mathcal{F}_N) = \frac{1}{N^2}\mathbf{a}_N^{\mathrm{T}}\boldsymbol{\Delta}_N\mathbf{a}_N = \frac{1}{N^2}\mathrm{Q}_{\boldsymbol{\Delta}_N}(\mathbf{a}_N) \leq \frac{\lambda_{\max}(\Delta_N)}{N}\frac{\|\mathbf{a}_N\|_2^2}{N}, \tag{5.9}$$

where $\|\mathbf{a}_N\|_2^2$ is the square of the L^2-norm of $\mathbf{a}_N$, $\|\mathbf{a}_N\|_2^2 = \sum_{k\in N} a_{Nk}^2$. The upper bound of $\mathbb{V}(Z_N \mid \mathcal{F}_N)$ is a function of the largest eigenvalue of $\boldsymbol{\Delta}_N$. Replacing $\lambda_{\max}(\Delta_N)$ by $K_N \geq \lambda_{\max}(\Delta_N)$ so

$$\mathbb{V}(Z_N \mid \mathcal{F}_N) \leq \frac{K_N}{N}\frac{\|\mathbf{a}_N\|^2}{N},$$

where K_N can be any of the following vector-induced matrix norms:

$$K_N = \begin{cases} \|\boldsymbol{\Delta}_N\|_1 = \max\limits_{l\in U}\sum\limits_{k=1}^{N}|\boldsymbol{\Delta}_{Nkl}| = \max\limits_{l\in U_N}\sum\limits_{k=1}^{N}|\pi_{Nkl} - \pi_{Nk}\pi_{Nl}| & \text{1- norm} \\ \|\boldsymbol{\Delta}_N\|_\infty = \max\limits_{k\in U_N}\sum\limits_{l=1}^{N}|\boldsymbol{\Delta}_{Nkl}| = \max\limits_{k\in U_N}\sum\limits_{l=1}^{N}|\pi_{Nkl} - \pi_{Nk}\pi_{Nl}| & \infty\text{-norm} \\ \|\boldsymbol{\Delta}_N\|_F = \left[\mathrm{tr}\left(\boldsymbol{\Delta}_N^{\mathrm{T}}\boldsymbol{\Delta}_N\right)\right]^{1/2} & \text{Frobenius norm} \end{cases}.$$

This upper bound depends on the values of the elements of $\boldsymbol{\Delta}_N$. As in the previous section, we can refine the upper bound for sample designs with random sample sizes since $\boldsymbol{\Delta}_N = \mathrm{diag}\left(\boldsymbol{\pi}_N \odot (1 - \boldsymbol{\pi}_N)\right)$ then

$$\lambda_{\max}\left(\Delta_N\right) = \max_{k \in U_N} \arg\left\{\left[\boldsymbol{\Delta}_{Nkk}\right]\right\} = \max_{k \in U_N} \arg\left\{\left[\pi_{Nk}\left(1-\pi_{Nk}\right)\right]\right\}. \tag{5.10}$$

Also, as in the previous section, the upper bound that does not depend on the eigenvalues or matrix norms is found by noting that $\lambda_{\max}\left(\Delta_N\right)$ is maximum when $\pi_k = \frac{1}{2}$, so

$$\mathbb{V}\left(Z_N \mid \mathcal{F}_N\right) \leq \frac{K_N}{N}\frac{\left\|\mathbf{a}_N\right\|^2}{N} = \mathcal{O}\left(\frac{1}{N}\right)\mathcal{O}(1) = \mathcal{O}\left(\frac{1}{N}\right), \tag{5.11}$$

where $K_N = 0.5$ after applying Slutsky's theorem and assuming that $\frac{\left\|\mathbf{a}_N\right\|_2^2}{N} = \mathcal{O}(1)$. An implicit assumption in (5.11) is that $\lambda_{\max}\left(\Delta_N\right) = \mathcal{O}(1)$ as $N \to \infty$. We explore situations where $\frac{\left\|\mathbf{a}_N\right\|_2^2}{N} \neq \mathcal{O}(1)$ in Section 5.10 by defining an explicit sequence $\left\{\mathbf{a}_N\right\}_{N=1}^{\infty}$.

5.8 The Horvitz-Thompson Estimator as a Linear Function of the Elements of the Random Vector S

The HT estimator of the population mean $\bar{Y} = \frac{1}{N}\mathbf{1}^{\mathrm{T}}\mathbf{y}$ is the linear function $Z(\mathbf{S})$ defined in Section 5.6 where $\mathbf{a} = \mathbf{d} \odot \mathbf{y}$, $\mathbf{d} = \mathbf{1} \oslash \boldsymbol{\pi} = \left[d_k\right] = \left[\pi_k^{-1}\right]$ for $k \in U$ and $\mathbf{y} \in \mathbb{R}^N$. The HT estimator of the mean $\bar{Y}$ is

$$\hat{\bar{Y}}_{HT} = Z(\mathbf{S}) = \frac{1}{N}(\mathbf{d} \odot \mathbf{y})^{\mathrm{T}} \mathbf{S}. \tag{5.12}$$

Let $\left\{\hat{\bar{Y}}_{HT,N}\right\}_{N=1}^{\infty}$ be the sequence of HT estimators defined in (5.12), then the expected value and variance are

$$\mathbb{E}\left(\hat{\bar{Y}}_{HT,N} \mid \mathcal{F}_N\right) = \frac{1}{N}\mathbf{1}_N^{\mathrm{T}} \mathbf{y}_N, \tag{5.13}$$

$$\mathbb{V}\left(\hat{\bar{Y}}_{HT,N} \mid \mathcal{F}_N\right) = \frac{1}{N^2}(\mathbf{d}_N \odot \mathbf{y}_N)^{\mathrm{T}} \Delta_N (\mathbf{d}_N \odot \mathbf{y}_N). \tag{5.14}$$

The conditions for the sequence of the estimators $\left\{\hat{\bar{Y}}_{HT,N}\right\}_{N=1}^{\infty}$ to be asymptotically unbiased and consistent depend on the sample design and the outcome (Särndal, Swensson, & Wretman, 1992). In other words, whether an estimator meets these conditions depend on the sequences $\{\mathbf{y}_N\}_{N=1}^{\infty}$, $\left\{\boldsymbol{\pi}_{N,k}\right\}_{N=1}^{\infty}$, and $\{\boldsymbol{\Delta}_N\}_{N=1}^{\infty}$. These cannot be set arbitrarily; for example, if the sequence of $\mathbf{S}_N$ is a valid sample design, then $\boldsymbol{\Delta}_N$ has to be a Hermitian matrix with the properties described in Sections 5.3.1 and 5.3.2. These additional conditions often are not fully explored in the current literature.

5.8.1 The Variance of the Horvitz-Thompson Estimator

To derive the variance of HT estimator, we reparametrize (5.12) using the variable $\breve{\mathbf{S}}$ defined as follows:

- Let $\breve{\mathbf{S}}: \mathbb{R}^N \to \mathbb{R}^N$ be a vector-to-vector valued function of $\mathbf{S}$ where $\breve{\mathbf{S}} = \mathbf{d} \odot \mathbf{S}$. The expected value of $\breve{\mathbf{S}}$ is

$$\mathbb{E}\left(\breve{\mathbf{S}} \mid \mathcal{F}\right) = \mathbf{d} \odot \mathbb{E}\left(\mathbf{S} \mid \mathcal{F}\right) = \mathbf{d} \odot \boldsymbol{\pi} = \mathbf{1}. \tag{5.15}$$

The covariance matrix of $\breve{\mathbf{S}}$, $\boldsymbol{\Delta}_{\breve{\mathbf{S}}} \in \mathbb{R}^{N \times N}$, is

$$\begin{aligned} \mathbb{V}\left(\breve{\mathbf{S}} \mid \mathcal{F}\right) &= \boldsymbol{\Delta}_{\breve{\mathbf{S}}} = \mathbf{d}^{\mathrm{T}} \mathbb{V}(\mathbf{S}) \mathbf{d} \\ &= \mathbf{d} \boldsymbol{\Delta} \mathbf{d}^{\mathrm{T}} = \boldsymbol{\Delta} \oslash \mathbf{d}^{\odot 2} = \left[\frac{d_k d_l}{d_{kl}} - 1 \right] \in \mathbb{R}^{N \times N}. \end{aligned} \tag{5.16}$$

The variance of the sequence of HT estimators, $\left\{ \hat{\bar{Y}}_{HT,N} \right\}_{N=1}^{\infty}$ is

$$\mathbb{V}\left(\bar{Y}_{HT,N} \mid \mathcal{F}_N\right) = \frac{1}{N^2} \mathbf{y}_N^{\mathrm{T}} \boldsymbol{\Delta}_{N\breve{\mathbf{S}}} \mathbf{y}_N = \frac{1}{N^2} \mathbf{Q}_{\boldsymbol{\Delta}_{N\breve{\mathbf{S}}}}\left(\mathbf{y}_N\right) \le \frac{\lambda_{\max}\left(\boldsymbol{\Delta}_{N\breve{\mathbf{S}}}\right)}{N} \frac{\|\mathbf{y}_N\|_2^2}{N}. \tag{5.17}$$

Its bound is a function of the largest eigenvalue, $\lambda_{\max}\left(\boldsymbol{\Delta}_{N\breve{\mathbf{S}}}\right)$, of the reparametrized covariance matrix $\boldsymbol{\Delta}_{N\breve{\mathbf{S}}}$. As in previous sections, we can refine the bound by replacing $\lambda_{\max}\left(\boldsymbol{\Delta}_{N\breve{\mathbf{S}}}\right)$ by $K_N \ge \lambda_{\max}\left(\boldsymbol{\Delta}_{N\breve{\mathbf{S}}}\right)$ using any of the matrix norms induced by the vector 1-norm, ∞-norm, or Frobenius norm as

$$K_N = \begin{cases} \left\| \boldsymbol{\Delta}_{N\breve{\mathbf{S}}} \right\|_1 = \max\limits_{l \in U_N} \sum\limits_{k=1}^{N} \left| \Delta_{N\breve{\mathbf{S}},kl} \right| = \max\limits_{l \in U_N} \sum\limits_{k=1}^{N} \left| d_{Nk} d_{Nl} d_{Nkl}^{-1} - 1 \right| & \text{1- norm} \\ \left\| \boldsymbol{\Delta}_{N\breve{\mathbf{S}}} \right\|_\infty = \max\limits_{k \in U_N} \sum\limits_{l=1}^{N} \left| \Delta_{N\breve{\mathbf{S}},kl} \right| = \max\limits_{k \in U_N} \sum\limits_{l=1}^{N} \left| d_{Nk} d_{Nl} d_{Nkl}^{-1} - 1 \right| & \infty\text{-norm} \\ \left\| \boldsymbol{\Delta}_{N\breve{\mathbf{S}}} \right\|_F = \left[\mathrm{tr}\left(\boldsymbol{\Delta}_{N\breve{\mathbf{S}}}^{\mathrm{T}} \boldsymbol{\Delta}_{N\breve{\mathbf{S}}} \right) \right]^{1/2} & \text{Frobenius norm} \end{cases}.$$

For random sample size designs, we can refine the value of K_N since $\mathbf{\Delta}_{N\breve{\mathbf{S}}} = \text{diag}\left(\mathbf{d}_N - 1\right)$. The value of K_N is

$$K_N = \underset{k \in U_N}{\arg\max} \left\{ \Delta_{N\breve{\mathbf{S}}} \right\} = d_{N\max} - 1 = \pi_{N\min} - 1.$$

Notice that effect of the weights $\mathbf{d}_N$ on $\mathbf{\Delta}_N$ reflected in $\mathbf{\Delta}_{N\breve{\mathbf{S}}}$. The bound of $\left\{ \hat{\bar{Y}}_{HT,N} \right\}_{N=1}^{\infty}$ is a function of the maximum sampling weight d_{Nk}, not the maximum π_{Nk} as in the estimator in Section 5.7. The bound of the variance of $\left\{ \hat{\bar{Y}}_{HT,N} \right\}_{N=1}^{\infty}$ is

$$\mathbb{V}\left(\bar{Y}_{HT,N} \mid \mathcal{F}_N \right) \leq \frac{K_N}{N} \frac{\left\| \mathbf{y}_N \right\|_2^2}{N} = \mathcal{O}\left(\frac{1}{N} \right) \mathcal{O}(1) = \mathcal{O}\left(\frac{1}{N} \right), \tag{5.18}$$

where $\left\| \mathbf{y}_N \right\|_2^2$ is the square of the Euclidian norm of $\mathbf{y}_N$, $\left\| \mathbf{y}_N \right\|_2^2 = \mathbf{y}_N^{\mathrm{T}} \mathbf{y}_N$. The order of the variance $\mathbb{V}\left(\bar{Y}_{HT,N} \mid \mathcal{F}_N \right)$ is $\mathcal{O}\left(N^{-1} \right)$ after using Slutsky's theorem. Two implicit assumptions in (5.18) are $K_N = \mathcal{O}(1)$ and $\frac{\left\| \mathbf{y}_N \right\|_2^2}{N} = \mathcal{O}(1)$.

Breidt & Opsomer (2017) lists two conditions for the consistency of the HT estimator:

D1: $$\limsup_{N \to \infty} \left\{ n \max_{k \neq l \in U_N} \left| \Delta_{N,kl} \right| \right\} < \infty$$

D2: $$\limsup_{N \to \infty} \frac{\sum_{k \in U_N} y_k^2}{N_N} < \infty$$

Note that $\frac{\|\mathbf{y}_N\|_2^2}{N_N} = \frac{\sum_{k=1}^{N_N} y_{Nk}^2}{N_N} = \frac{y_{N1}^2 + ... + y_{NN_N}^2}{N_N} = \mathcal{O}(1)$ in (5.18) means that for the finite population second moment for y, there is a non-zero constant c such that

$$\frac{\sum_{k=1}^{N_N} y_{Nk}^2}{N_N} \to c < \infty, \tag{5.19}$$

as the population size increases. This is condition D2. To understand condition D1, we use the bound proposed by Breidt & Opsomer (2017) for the variance of the HT estimator in (5.20):

$$\mathbb{V}\left(\hat{Y}_{HT}\right) \leq \frac{1}{N\lambda_1} \sum_{k=1}^{N_N} \frac{y_{Nk}^2}{N} + \frac{\max_{k \neq l \in U_N} \left|\Delta_{N,kl}\right|}{\lambda_1^2} \left(\sum_{k=1}^{N_N} \frac{|y_k|}{N} \right)^2, \tag{5.20}$$

where $\min_{k \in U} \{\pi_k\} \geq \lambda_1 > 0$. Letting $\min_{k \in U} \{\pi_k\} = \lambda_1$ and defining $d = \frac{1}{\min \arg_{k \in U} \{\pi_k\}}$ then $d = \max \arg_{k \in U} \left\{ \frac{1}{\pi_k} \right\}$, that is, d is the maximum weight. Replacing λ_1 by d in (5.20) and after simplification using the fact that $\sum_{k=1}^{N} \pi_k = n$ and $f = \frac{n}{N}$, we obtain

$$\mathbb{V}\left(\hat{Y}_{HT}\right) = \mathcal{O}\left(\frac{1}{n}\right). \tag{5.21}$$

which converges to zero because $\max_{k \neq l, k, l \in U_N} \left|\Delta_{N,kl}\right| \to 0$ as $N \to \infty$. This result is based on the fact that draws from the sample tend to become independent, (e.g., $\pi_{kl} - \pi_k \pi_l \to 0$ for $k \neq l \in U$) as the population and sample sizes go to infinity.

Although both formulas give the same solution, (5.18) is easier to derive and interpret.

5.8.2 The Variance Estimator of the Horvitz-Thompson Estimator

The variance estimator of the HT estimator of the mean $\bar{Y}$ is derived from (5.16) after replacing $\boldsymbol{\Delta}$ by $\tilde{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$ as

$$\hat{\mathbb{V}}\left(\bar{Y}_{HT} \mid \mathcal{F}\right) = \frac{1}{N^2}\left(\mathbf{y} \odot \mathbf{d} \odot \mathbf{S}\right)^{\mathrm{T}} \breve{\boldsymbol{\Delta}}\left(\mathbf{y} \odot \mathbf{d} \odot \mathbf{S}\right). \tag{5.22}$$

Reparametrize $\hat{\mathbb{V}}\left(\bar{Y}_{HT} \mid \mathcal{F}\right)$ as a sum of the new variable $\psi_{kl} = \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\Delta_{kl}$ expanded by π_{kl}, similar to an HT estimator as

$$\hat{\mathbb{V}}\left(\bar{Y}_{HT} \mid \mathcal{F}\right) = \frac{1}{N^2}\sum_{k \in U}\sum_{l \in U}\frac{\psi_{kl}}{\pi_{kl}}. \tag{5.23}$$

Continue reparametrizing (5.23) using the following variables

- $\boldsymbol{\psi} \in \mathbb{R}^{N \times N}$ where $\boldsymbol{\psi} = \left(\mathbf{y} \oslash \boldsymbol{\pi}\right)^{\mathrm{T}} \boldsymbol{\Delta}\left(\mathbf{y} \oslash \boldsymbol{\pi}\right)$.
- $\mathbf{S}_2 \in \mathbb{R}^{N \times N}$, a matrix with the sample membership indicators of the 2-tuples (k,l) where $\mathbb{E}\left(\mathbf{S}_2\right) = \boldsymbol{\Pi}$, the matrix with the second order probability of inclusion π_{kl}.

- $\boldsymbol{\Delta}_{\mathbf{S}_2} \in \mathbb{R}^{N^2 \times N^2}$, the covariance matrix of $\mathbf{S}_2$ where $\boldsymbol{\Delta}_{\mathbf{S}_2} = \left[\pi_{klmn} - \pi_{kl}\pi_{mn}\right]$ and π_{klmn} is the fourth order inclusion probability of the 4-tuples (k,l,m,n).

- To avoid tensor notation, we vectorize $\boldsymbol{\psi}$ and $\boldsymbol{\Pi}$ as $\text{vec}(\boldsymbol{\psi}) \in \mathbb{R}^{N^2}$, $\text{vec}\left(\boldsymbol{\Pi}^{\odot -1}\right) \in \mathbb{R}^{N^2}$ (Magnus & Neudecker, 1999). The expression of $\hat{\mathbb{V}}\left(\bar{Y}_{HT} \mid \mathcal{F}\right)$ with the reparametrized variables is

$$\hat{\mathbb{V}}\left(\bar{Y}_{HT} \mid \mathcal{F}\right) = \frac{1}{N^2} \text{vec}(\boldsymbol{\psi})^{\mathrm{T}} \text{vec}\left(\boldsymbol{\Pi}^{\odot -1} \odot \mathbf{S}_2\right). \tag{5.24}$$

The expected value is

$$\begin{aligned} \mathbb{E}\left(\hat{\mathbb{V}}\left(\bar{Y}_{HT}\right) \mid \mathcal{F}\right) &= \frac{1}{N^2} \text{vec}(\boldsymbol{\psi})^{\mathrm{T}} \mathbb{E}\left(\text{vec}\left(\boldsymbol{\Pi}^{\odot -1} \odot \mathbf{S}_2\right)\right) \\ &= \frac{1}{N^2} \text{vec}(\boldsymbol{\psi})^{\mathrm{T}} \text{vec}\left(\boldsymbol{\Pi}^{\odot -1} \odot \mathbb{E}(\mathbf{S}_2)\right) \\ &= \frac{1}{N^2} \text{vec}(\boldsymbol{\psi})^{\mathrm{T}} \text{vec}\left(\boldsymbol{\Pi}^{\odot -1} \odot \boldsymbol{\Pi}\right) \\ &= \frac{1}{N^2} \text{vec}(\boldsymbol{\psi})^{\mathrm{T}} \mathbf{1}_{N^2} = \mathbb{V}\left(\bar{Y}_{HT}\right) \end{aligned}, \tag{5.25}$$

therefore, $\hat{\mathbb{V}}\left(\bar{Y}_{HT} \mid \mathcal{F}\right)$ is an unbiased estimator of $\mathbb{V}\left(\bar{Y}_{HT} \mid \mathcal{F}\right)$.

To study the limiting distribution and bounds of the estimator $\hat{\mathbb{V}}\left(\bar{Y}_{HT} \mid \mathcal{F}\right)$ as $N, n \to \infty$, we derive the expression of $\mathbb{V}\left(\hat{\mathbb{V}}\left(\bar{Y}_{HT}\right) \mid \mathcal{F}\right)$ following the same procedures from the previous sections.

$$
\begin{aligned}
\mathbb{V}\left(\hat{\mathbb{V}}\left(\bar{Y}_{HT}\right)|\mathcal{F}\right) &= \frac{1}{N^4}\mathbb{V}\left(\text{vec}\left(\boldsymbol{\psi}\oslash\boldsymbol{\Pi}\right)^{\mathrm{T}}\text{vec}\left(\mathbf{S}_2\right)\right) = \\
&= \frac{1}{N^4}\text{vec}\left(\boldsymbol{\psi}\oslash\boldsymbol{\Pi}\right)^{\mathrm{T}}\mathbb{V}\left(\text{vec}\left(\mathbf{S}_2\right)\right)\text{vec}\left(\boldsymbol{\psi}\oslash\boldsymbol{\Pi}\right) \\
&= \frac{1}{N^4}\text{vec}\left(\boldsymbol{\psi}\oslash\boldsymbol{\Pi}\right)^{\mathrm{T}}\boldsymbol{\Delta}_{\mathbf{S}_2}\text{vec}\left(\boldsymbol{\psi}\oslash\boldsymbol{\Pi}\right) \\
&= \frac{1}{N^4}Q_{\boldsymbol{\Delta}_{\mathbf{S}_2}}\left(\text{vec}\left(\boldsymbol{\psi}\oslash\boldsymbol{\Pi}\right)\right) \\
&\leq \frac{\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{S}_2}\right)}{N^3}\frac{\left\|\mathbf{y}\odot\mathbf{y}\right\|_2^2}{N}
\end{aligned}
,
$$

where $\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{S}_2}\right)$ is the largest eigenvalue of the matrix

$$
\boldsymbol{\Sigma}_{\mathbf{S}_2} = \boldsymbol{\Delta}_{\mathbf{S}_2}\odot\boldsymbol{\Delta}^{\odot 2}\oslash\boldsymbol{\pi}^{\odot 2}\oslash\boldsymbol{\Pi}^{\odot 2},
$$

with the element $\boldsymbol{\Sigma}_{klmn,\mathbf{S}_2} = \frac{\left(\pi_{kl}-\pi_k\pi_l\right)^2}{\pi_k^2\pi_l^2}\frac{\left(\pi_{klmn}-\pi_{kl}\pi_{mn}\right)}{\pi_{kl}^2}$.

An upper bound $K \geq \lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbf{S}_2}\right)$ is obtained using the vector induced matrix norms in $\boldsymbol{\Sigma}_{\mathbf{S}_2}$ as

$$
K = \begin{cases}
\left\|\boldsymbol{\Sigma}_{\mathbf{S}_2}\right\|_1 = \max\limits_{l\in U}\sum\limits_{k=1}^{N}\left|\boldsymbol{\Sigma}_{\mathbf{S}_2,kl}\right| & \text{1- norm} \\
\left\|\boldsymbol{\Sigma}_{\mathbf{S}_2}\right\|_\infty = \max\limits_{k\in U}\sum\limits_{l=1}^{N}\left|\boldsymbol{\Sigma}_{\mathbf{S}_2,kl}\right| & \infty\text{-norm} \\
\left\|\boldsymbol{\Sigma}_{\mathbf{S}_2}\right\|_F = \left[\text{tr}\left(\boldsymbol{\Sigma}_{\mathbf{S}_2,kl}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{S}_2,kl}\right)\right]^{1/2} & \text{Frobenius norm}
\end{cases}
.
$$

The main difficulty of identifying an upper bound for K is that it requires examining the elements of $\boldsymbol{\Sigma}_{\mathbf{S}_2}$ where the third and fourth order $\pi_{klm}\pi_{klmn}$ of inclusion

probabilities (π_{klm} and π_{klmn}) are not available or difficult to compute for some complex designs.

On the other hand, for random sample size designs, we can refine the value of K since $\boldsymbol{\Sigma}_{\mathbf{S}_2}$ is a diagonal matrix where $\boldsymbol{\Sigma}_{\mathbf{S}_2} = \left[\left(d_k - 1\right)^3\right] = \left[\left(\pi_k^{-1} - 1\right)^3\right]$. K is the maximum sampling weight which is equivalent to the smallest π_k. Assuming that $\frac{\|\mathbf{y} \odot \mathbf{y}\|_2^2}{N} = \mathcal{O}(1)$ then, after using Slutsky's theorem,

$$\mathbb{V}\left(\hat{\mathbb{V}}\left(\bar{Y}_{HT}\right) | \mathcal{F}\right) \leq \frac{K}{N^3} \frac{\|\mathbf{y} \odot \mathbf{y}\|_2^2}{N} = \mathcal{O}\left(\frac{1}{N^3}\right) \mathcal{O}(1) = \mathcal{O}\left(\frac{1}{N^3}\right). \tag{5.26}$$

$\hat{\mathbb{V}}\left(\bar{Y}_{HT}\right)$ is bounded in probability and $\lim_{N \to \infty} \hat{\mathbb{V}}\left(\bar{Y}_{HT,N}\right) = \lim_{N \to \infty} \mathbb{V}\left(\bar{Y}_{HT,N}\right) = 0$. The expression in (5.26) implicitly assumes that $\frac{\|\mathbf{y} \odot \mathbf{y}\|_2^2}{N}$ is $\mathcal{O}(1)$ which can be written as

$$\frac{\|\mathbf{y} \odot \mathbf{y}\|_2^2}{N} = \frac{\left\|\mathbf{y}^{\odot 2}\right\|_2^2}{N} = \frac{\sum_{k=1}^{N}\left(y_k^2\right)^2}{N} = \frac{\sum_{k=1}^{N} y_k^4}{N} = \mathcal{O}(1), \tag{5.27}$$

which is the fourth population moment of y. Equation (5.26) is condition D4 in Breidt & Opsomer (2017). Condition D2 is $\min_{k,l \in U_N} \{\pi_{Nkl}\} \geq \lambda > 0$ which we have already covered since, in order to produce $\hat{\mathbb{V}}\left(\bar{Y}_{HT} | \mathcal{F}\right)$, we divide by $\boldsymbol{\Delta}$ by $\boldsymbol{\Pi}$ which is defined if $\min_{k,l \in U} \arg\left\{\left|\pi_{kl}\right|\right\} > 0$. The result in (5.26) is found in the literature.

We illustrate the speed of convergence varies and we can even find situations where $\hat{\mathbb{V}}\left(\bar{Y}_{HT} \mid \mathcal{F}\right)$ will not become zero as $N \to \infty$. Substitute x_k^2 by y_k^4 in $\left\|\mathbf{y} \odot \mathbf{y}\right\|_2^2$, then an upper bound of $\hat{\mathbb{V}}\left(\bar{Y}_{HT,N}\right)$, in terms of the population mean $\bar{Y}_N$, is $\hat{\mathbb{V}}\left(\bar{Y}_{HT,N}\right) \leq K\bar{X}_N^2 = K_N N^2 \bar{Y}_N^4$. If we define $\left\{\mathbf{y}_N\right\}_{N=1}^{\infty}$ as a sequence of real constants, $\mathbf{y}_N \in \mathbb{R}^N$ where $\bar{Y}_N = \mathcal{O}\left(N^P\right)$, then the value of p such as $\hat{\mathbb{V}}\left(\bar{Y}_{HT,N}\right)$ does not converge, e.g., $\hat{\mathbb{V}}\left(\bar{Y}_{HT,N}\right) \geq \mathcal{O}_p(1)$, is $p \geq -\frac{1}{2}$. If $-\frac{3}{4} < p < -\frac{1}{2}$ then $\mathbb{V}\left(\bar{Y}_{HT,N}\right)$ converges at a slower rate than $\mathcal{O}_p\left(N^{-1}\right)$; if $p < -\frac{3}{4}$, $\mathbb{V}\left(\bar{Y}_{HT,N}\right)$ converges at a faster rate than $\mathcal{O}_p\left(N^{-1}\right)$.

5.8.3 The Central Limit Theorem and the Horvitz-Thompson Estimator

Deriving the asymptotic normality of a design-based estimator is a difficult topic. The Central Limit Theorem (CLT) for finite populations has only been rigorously justified for some designs (Cardot, Degras, & Josserand, 2013). Proof for equal probability sampling is found in Madow (1948), Erdös & Rényi (1959), and Hájek (1960) while Hájek (1964) proved the theorem for rejective Poisson sampling with varying probabilities and Scott & Wu (1981) for the ratio and regression estimators under simple random sampling. In general, the finite population CLT proofs are technically difficult and omitted in most textbooks. Using the multivariate approach for the

random vector **S** and the fixed finite population **y** provides an alternative approach for proving the theorem for some designs.

Consider all designs where the sampling units are independently drawn without replacement, $\pi_{kl} = \pi_k \pi_l$ for $k \neq l \in U$. Examples of these designs are Bernoulli and Poisson. Using the re-parametrization described in Section 5.8.1, then $\hat{\bar{Y}}_{HT} = \bar{\breve{S}}$ with $\mathbb{E}\left(\breve{S}_k \mid \mathcal{F}\right) = y_k$, and $\mathbb{V}\left(\breve{S}_k \mid \mathcal{F}\right) = y_k^2 \left(d_k - 1\right)$. By the Lindenberg, Lévy, and Feller version of the CLT for independent random variables with different means and variances (see Theorem 6.1 and Corollary 6.3 in Polansky 2011), the sequence of estimators $\left\{Z_N\right\}_{N=1}^{\infty}$, $Z_N = N\tau_N^{-1}\left(\hat{\bar{Y}}_{HT,N} - \bar{Y}_N\right)$ and $\tau_N = \sum_{k \in U_N} \mathbb{V}\left(\breve{S}_{Nk} \mid \mathcal{F}\right)$ has a limit distribution $\mathcal{N}\left(0,1\right)$ if $\sum_{k \in U_N} \mathbb{E}\left(\left|S_k - y_k\right|^{\eta}\right) = o\left(\tau_N^{\eta}\right)$ for some $\eta > 2$.

For other designs, where $\breve{S}_k$ and $\breve{S}_l$ are correlated, the Lindenberg-Lévy-Feller CLT assumption of independence may be weakened. For example, if we redefine the sequence $\left\{\hat{\bar{Y}}_{HT,N}\right\}_{N=0}^{\infty} = \left\{\bar{\breve{S}}_N\right\}_{N=0}^{\infty}$ as a sequence of dependent and correlated random variables and we assume the following conditions hold:

$$\begin{aligned} \mathbb{E}\left(\breve{S}_{N,k} \mid \mathcal{F}_N\right) &= \mu_{N,k} < \infty, \\ \mathbb{V}\left(\breve{S}_N \mid \mathcal{F}_N\right) &= \tau_{N,k} < \infty, \text{ and} \\ \lim_{N \to 0} N\mathbb{V}\left(\bar{\breve{S}}_N \mid \mathcal{F}_N\right) &= \tau_N \in \left(0, \infty\right), \end{aligned} \tag{5.28}$$

where $\mu_{N,k} = \frac{1}{N}\sum_{k\in U}\pi_k$, and $\tau_N = \sum_{k\in U_N}\mathbb{V}\left(\breve{S}_{Nk} \mid \mathcal{F}_N\right) + \sum\sum_{k\neq l\in U_N}\mathbb{C}\left(\breve{S}_{Nk}, \breve{S}_{Nl} \mid \mathcal{F}_N\right)$ for $N \to \infty$. Several mild technical conditions but different from author to author need to be imposed beyond those for Lindeberg-Lévy-Feller to derive Central Limit Theorems for dependent correlated sequences. Most authors claim a limiting normal distribution by appealing to the specific version of the central limit theorem. For example, Breidt, Opsomer, & Sanchez-Borrego, (2016) claim normality after invoking Lyapunov's version of the central limit theorem.

5.8.4 The Design Consistency of the Horvitz-Thompson Estimator

There are different ways to establish consistency of a sequence of estimators $\left\{\hat{\theta}_N\right\}_{N=1}^{\infty}$. For example, Lehmann (1999) gives a sufficient condition for an estimator to be consistent when the sequence of estimators converge to a constant in quadratic mean. This condition is demonstrated for the HT estimator using the same reparameterization of (5.12) with $\breve{\mathbf{S}} = \mathbf{d} \odot \mathbf{y} \odot \mathbf{S}$. The expected value and variance of the HT estimator, $\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{k\in U}\breve{S}_i$, is

$$\mathbb{E}\left(\hat{\bar{Y}}_{HT} \mid \mathcal{F}\right) = \bar{Y}, \text{ and}$$

$$\mathbb{V}\left(\hat{\bar{Y}}_{HT} \mid \mathcal{F}\right) = \frac{1}{N^2}\left(\sum_{k\in U}\sum_{l\in U}\Delta_{\breve{\mathbf{S}},kl}\right) = \frac{1}{N^2}\left(\sum_{k\in U}\mathbb{V}\left(\breve{S}_k \mid \mathcal{F}\right) + \sum\sum_{k\neq l\in U}\mathbb{C}\left(\breve{S}_k, \breve{S}_l \mid \mathcal{F}\right)\right), \quad (5.29)$$

where the terms $\sum_{k\neq l\in U}\sum \mathbb{C}\left(\breve{S}_k,\breve{S}_l \mid \mathcal{F}\right)$ are not zero. Let $\left\{\hat{\bar{Y}}_{HT,N}\right\}_{N=1}^{\infty}$ be the sequence of HT estimators, where $\hat{\bar{Y}}_{HT,N}=\frac{1}{N_N}\sum_{k\in U_N}\breve{S}_{k,N}$, then

$$\Pr\left[\left(\hat{\bar{Y}}_{HT,N}-\hat{\bar{Y}}_N\right)^2\geq\varepsilon_N^2\right]\leq\frac{\mathbb{V}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}_N\right)}{\varepsilon_N^2}=\frac{1}{N^2}\frac{\sum_{k\in U_N}\sum_{k\in U_N}\Delta_{Nkl}}{\varepsilon_N^2}$$
$$\leq\frac{K_N}{\varepsilon_N^2 N_N}\frac{\left\|\mathbf{y}_N\right\|_2^2}{N_N} \qquad (5.30)$$

Since (5.30) holds for any $\varepsilon_N>0$ then the sequence of HT estimators $\left\{\hat{\bar{Y}}_{HT,N}\right\}_{N=1}^{\infty}$ is consistent for $\bar{Y}_N$. Note that this condition holds for both random sample size and fixed sample size single stage designs.

For our discussion, to prove that a sequence of estimators $\left\{\hat{\theta}_N\right\}_{N=1}^{\infty}$ is design consistent of the population characteristic θ_N, we use two sufficient conditions to establish design consistency (Remark 5.3.1 and Exercise 5.18 in Särndal, Swensson, & Wretman 1992):

(a.) The sequence of estimators $\left\{\hat{\theta}_N\right\}_{N=1}^{\infty}$ from sequences of sample sizes $\left\{n_N\right\}_{N=1}^{\infty}$ drawn using sample designs $\left\{p_N\left(A_N=a_N\right)\right\}_{N=1}^{\infty}$ from the sequence of populations $\left\{\mathcal{F}_N\right\}_{N=1}^{\infty}$ of increasing sample sizes $\left\{N_N\right\}_{N=1}^{\infty}$, is asymptotically unbiased for a population characteristic θ_N, that is

$$\lim_{N\to\infty}\left[\mathbb{E}\left(\hat{\theta}_N\right)-\theta_N\right]=0\,. \tag{5.31}$$

(b.) The variance of the sequence of estimators $\left\{\hat{\theta}_N\right\}_{N=1}^{\infty}$ goes to zero as the sample and population sizes go to infinity (e.g., $\mathbb{V}\left(\hat{\theta}_N \mid \mathcal{F}\right)<\infty$ and $\lim_{N\to\infty}\mathbb{V}\left(\hat{\theta}_N \mid \mathcal{F}\right)=0$).

The design consistency of the HT estimator of the mean is proven using the results (5.13) and (5.18), the sequence of HT estimators $\left\{\hat{\bar{Y}}_{HT,N}\right\}_{N=1}^{\infty}$ is design consistent for $\bar{Y}_N$.

5.8.5 The Confidence Intervals and the Horvitz-Thompson Estimator

In this section, we derive the asymptotic properties of the confidence intervals (CI) of the HT estimator of the mean. Confidence intervals are created by identifying a function of the observed sample data that produces an interval or region containing the true parameter value with a probability α (e. g., 100α % or confidence coefficient) that is specified before selecting the sample (Polansky, 2011). CIs are created by inverting a statistical hypothesis test or a pivotal quantity defined as a function of the data and the unknown parameter θ, whose distribution does not depend on θ or any other unknown parameter (Casella & Berger, 2002).

Since we use the results of the CLT, the confidence intervals also refer to a sequence of random variables. For example, for the sequence for the HT estimator $\left\{\hat{\bar{Y}}_{HT,N}\right\}_{N=1}^{\infty}$ of $\bar{Y}_N$, the function $Z=\dfrac{\hat{\bar{Y}}_{HT}-\bar{Y}}{\sqrt{\mathbb{V}\left(\hat{\bar{Y}}_{HT}\mid\mathcal{F}\right)}}$, the sequence is $\{Z_N\}_{N=1}^{\infty}$, where $Z_N=\dfrac{\hat{\bar{Y}}_{HT,N}-\bar{Y}_N}{\sqrt{\mathbb{V}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}\right)}}$. The limiting distribution of $\{Z_N\}_{N=1}^{\infty}$ is $\mathcal{N}(0,1)$ as $N\to\infty$.

For confidence intervals, we define the sequence $\left\{C_N(\alpha\mid\mathcal{F}_N)\right\}_{N=1}^{\infty}$ in terms of the upper and lower limits as

$$C_N(\alpha\mid\mathcal{F}_N)= \left[\bar{Y}_{HT,N}+\sqrt{\mathbb{V}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}_N\right)}z_{(1-\alpha)/2},\bar{Y}_{HT,N}+\sqrt{\mathbb{V}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}_N\right)}z_{(1+\alpha)/2}\right]. \quad (5.32)$$

When $\mathbb{V}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}_N\right)$ is not known we replace it by $\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}_N\right)$, and the revised sequence $\left\{Z_N^*\right\}_{N=1}^{\infty}$ where $Z_N^*=\dfrac{\hat{\bar{Y}}_{HT,N}-\bar{Y}_N}{\sqrt{\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}_N\right)}}$ converges to a normal distribution. This result follows because the sequence $\left\{\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}_N\right)\right\}_{N=1}^{\infty}$ is a consistent estimator of $\mathbb{V}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}_N\right)$; using the theorem for functions of consistent estimators, $\sqrt{\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{HT,N}\mid\mathcal{F}_N\right)}\xrightarrow{P}\sqrt{\hat{\mathbb{V}}\left(\bar{Y}_{HT,N}\mid\mathcal{F}_N\right)}$. Combining all these results,

$$\begin{aligned}
&\lim_{N\to\infty} \Pr\left[\overline{Y}_N \in \hat{C}_N\left(\alpha \mid \mathcal{F}_N\right)\right] \\
&= \lim_{N\to\infty} \Pr\left[\sqrt{\mathbb{V}\left(\hat{\overline{Y}}_{HT,N} \mid \mathcal{F}_N\right)} z_{(1-\alpha)/2} \le \overline{Y}_N - \overline{Y}_{HT,N} \le \sqrt{\mathbb{V}\left(\hat{\overline{Y}}_{HT,N} \mid \mathcal{F}_N\right)} z_{(1+\alpha)/2}\right]. \qquad (5.33) \\
&= \alpha
\end{aligned}$$

In other words, the sequence of upper and lower limits of the confidence intervals $\hat{C}_N\left(\alpha \mid \mathcal{F}\right)$ are asymptotically accurate.

5.9 Properties of Estimators as Nonlinear Functions of the Elements of S

All PA estimators are functions of $\mathbf{S}$, which is a consistent estimator of $\boldsymbol{\pi}$. We can derive the asymptotic properties of new estimators under regularity conditions that depend on the type of function. In Section 5.8 we derive the large sample properties of the HT estimator which is a linear function of the $S_k \in \mathbf{S}$.

For estimators such as the HJ and ratio estimators, the function is nonlinear; that is, the estimator is a ratio of linear combinations of $S_k \in \mathbf{S}$. For this type of estimators, the variance is derived using the linear approximation of the nonlinear function using the first two terms of the Taylor Series (TS) expansion.

The PA estimator of the mean of the population characteristic θ is defined as $\mathrm{f}: \mathbb{R}^N \mapsto \mathbb{R}$, the vector-to-scalar valued function twice differentiable, where

$$\hat{\overline{\theta}}\left(\mathbf{S}\right) = \frac{1}{N}\mathbf{d}^{\mathrm{T}}\left(\mathbf{S} \odot \mathrm{f}\left(\mathbf{S}\right)\right). \qquad (5.34)$$

The TS approximation of $\hat{\bar{\theta}}(\mathbf{S})$ evaluated at the point $\mathbf{S}=\boldsymbol{\pi}$. is

$$\hat{\bar{\theta}}(\mathbf{S}) = \frac{1}{N}\mathbf{d}^{\mathrm{T}}\left(\mathbf{S}\odot \mathrm{f}(\mathbf{S})\right)\Big|_{\mathbf{S}=\boldsymbol{\pi}} + \frac{1}{N}\frac{\partial \mathbf{d}^{\mathrm{T}}\left(\mathbf{S}\odot \mathrm{f}(\mathbf{S})\right)}{\partial \mathbf{S}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}} (\mathbf{S}-\boldsymbol{\pi}) + \frac{1}{N}\mathcal{O}_p\left(\left\|\mathbf{S}^*-\boldsymbol{\pi}\right\|_2^2\right), \tag{5.35}$$

where $\left\|\mathbf{S}^*-\boldsymbol{\pi}\right\| \le \left\|\mathbf{S}-\boldsymbol{\pi}\right\|$. Thus, the expected value $\mathbb{E}\left(\hat{\bar{\theta}}(\mathbf{S})\right)$ is

$$\mathbb{E}\left(\hat{\bar{\theta}}(\mathbf{S})\,|\,\mathcal{F}\right) = \frac{1}{N}\mathbf{1}^{\mathrm{T}}\mathrm{f}(\boldsymbol{\pi}) + \mathcal{O}\left(N^{-1}\right), \tag{5.36}$$

because $\mathbb{E}(\mathbf{S}-\boldsymbol{\pi}\,|\,\mathcal{F})=\mathbf{0}$ and $\mathbb{E}\left(\left\|\mathbf{S}^*-\boldsymbol{\pi}\right\|_2^2\,|\,\mathcal{F}\right) \le \frac{C}{N}\bar{\pi}$ for a constant C.

The variance $\mathbb{V}\left(\hat{\bar{\theta}}(\mathbf{S})\,|\,\mathcal{F}\right)$ is

$$\mathbb{V}\left(\hat{\bar{\theta}}(\mathbf{S})\,|\,\mathcal{F}\right) = \frac{1}{N^2}\left(\left(\frac{\partial \mathbf{d}^{\mathrm{T}}\left(\mathbf{S}\odot \mathrm{f}(\mathbf{S})\right)}{\partial \mathbf{S}}\right)\Big|_{\mathbf{S}=\boldsymbol{\pi}}\right)^{\mathrm{T}} \boldsymbol{\Delta}\left(\frac{\partial \mathbf{d}^{\mathrm{T}}\left(\mathbf{S}\odot \mathrm{f}(\mathbf{S})\right)}{\partial \mathbf{S}}\right)\Big|_{\mathbf{S}=\boldsymbol{\pi}} + \mathcal{O}\left(\frac{1}{N}\right). \tag{5.37}$$

The approximate variance of $\hat{\bar{\theta}}(\mathbf{S})$ is

$$\mathbb{AV}\left(\hat{\bar{\theta}}(\mathbf{S})\,|\,\mathcal{F}\right) = \frac{1}{N^2}\left(\left(\frac{\partial \mathbf{d}^{\mathrm{T}}\left(\mathbf{S}\odot \mathrm{f}(\mathbf{S})\right)}{\partial \mathbf{S}}\right)\Big|_{\mathbf{S}=\boldsymbol{\pi}}\right)^{\mathrm{T}} \boldsymbol{\Delta}\left(\frac{\partial \mathbf{d}^{\mathrm{T}}\left(\mathbf{S}\odot \mathrm{f}(\mathbf{S})\right)}{\partial \mathbf{S}}\right)\Big|_{\mathbf{S}=\boldsymbol{\pi}}. \tag{5.38}$$

We now derive the regularity conditions that will permit us to establish the large-sample properties of the PA estimator based on the function $\mathrm{f}(\mathbf{S})$ of the discrete random vector $\mathbf{S}$. We do not include any regularity conditions for the existence and

uniqueness of maximum likelihood estimators and pseudo-maximum likelihood estimators which are part of the PA framework (see Definitions 1.7 and 1.10).

Let $\mathrm{f}(\mathbf{S})$ be the PA estimator for population characteristic θ, define the following

1. Let $\left\{\mathrm{f}(\mathbf{S}_N)\right\}_{N=1}^{\infty}$ be a sequence of estimators defined by $\mathrm{f}(\mathbf{S}_N)$ for a sequence of nested finite populations $\{\mathcal{F}_N\}_{N=1}^{\infty}$ such as $\mathcal{F}_N \subset \mathcal{F}_{N+1}$ for $N \in \mathbb{N}$ with increasing population size where each element of the population is identified by their labels $U_N \in \{1,...,N_N\}$.

2. Each population $\mathcal{F}_{N'} \in \{\mathcal{F}_N\}_{N=1}^{\infty}$ in the sequence $\mathcal{F}_{N'} = (\mathbf{y}_{N'}, \mathbf{x}_{N'}) \in \mathbb{R}^{N' \times (P+1)}$ consists of a vector with the population characteristic of interest $y_{N'} \in \mathbb{R}^{N' \times 1}$ and a matrix $\mathbf{x}_{N'} \in \mathbb{R}^{N' \times P}$ with *P*-auxiliary variables.

3. Let $\{\mathbf{S}_N\}_{N=1}^{\infty}$ be a sequence or random vectors with the sample membership indicators associated with the sequence of populations $\{\mathcal{F}_N\}_{N=1}^{\infty}$.

4. Each k sample membership indicator $S_{k,N'} \in \mathbf{S}_{N'}$ is associated with the k element of the finite population $\mathcal{F}_{kN'} = (y_{kN'}, \mathbf{x}_{kN'})$ for $k \in U_{N'}$ for each $\mathbf{S}_{N'} \in \{\mathbf{S}_N\}_{N=1}^{\infty}$ and $\mathcal{F}_{N'} \in \{\mathcal{F}_N\}_{N=1}^{\infty}$.

5. For each $\mathbf{S}_{N'} \in \{\mathbf{S}_N\}_{N=1}^{\infty}$, the expected value and variance-covariance of $\mathbf{S}_{N'}$, $\mathbb{E}(\mathbf{S}_{N'}) = \boldsymbol{\pi}_{N'}$ and $\mathbb{C}(\mathbf{S}_{N'}) = \boldsymbol{\Delta}_{N'}$, uniquely define sample design $p_{N'}(A_{N'} = a_{N'})$ for the population N' in the sequence of sample designs $\{p_N(A_N = a_N)\}_{N=1}^{\infty}$ associated with $\{\mathcal{F}_N\}_{N=1}^{\infty}$.

6. For each $\mathbf{S}_{N'} \in \{\mathbf{S}_N\}_{N=1}^{\infty}$ in $\mathcal{F}_{N'} \in \{\mathcal{F}_N\}_{N=1}^{\infty}$, the sample design is measurable, that is $\pi_{kN'} > 0$ for all $k \in U_N'$, and $\pi_{klN'} > 0$ for all $k \neq l \in U_N'$ $\mathbb{E}(\mathbf{S}_{N'}) = \boldsymbol{\pi}_{N'}$ and $\mathbb{C}(\mathbf{S}_{N'}) = \boldsymbol{\Delta}_{N'}$.

7. For each $\mathbf{S}_{N'} \in \{\mathbf{S}_N\}_{N=1}^{\infty}$ in $\mathcal{F}_{N'} \in \{\mathcal{F}_N\}_{N=1}^{\infty}$, the sample size drawn from the population $\mathcal{F}_{N'}$ is $n_{N'} = \mathbf{1}_{N'}^1 \boldsymbol{\pi}_{N'} = \sum_{k \in N'} \pi_{kN'}$ for fixed sample size designs, or the expected sample size is $\mathbb{E}(n_N) = \mathbf{1}_N^1 \boldsymbol{\pi}_N$. We assume that $\lim_{N \to \infty} \frac{n_{N'}}{N_N} = f \in (0,1)$, that is as the population size goes to infinity, the ratio converges to the overall sampling rate bounded and away from 0 or 1.[23]

[23] Note that we do not assume that the sample size goes to infinity. The increasing population size affects the sample design $\mathbf{S}$ which affects $\mathbb{E}(\mathbf{S}_{N'}) = \boldsymbol{\pi}_{N'}$. In other words, the sample size n cannot set separately from $N \to \infty$.

8. For each $\mathbf{y}_{N'} \in \{\mathcal{F}_N\}_{N=1}^{\infty}$ in the sequence of populations, $\mathcal{F}_{N'} \in \{\mathcal{F}_N\}_{N=1}^{\infty}$, the Euclidian norm is bounded, $\frac{\|\mathbf{y}_N\|_2^2}{N_N} = \mathcal{O}(1)$, as $N \to \infty$ (for consistency of the PA estimator $\mathrm{f}(\mathbf{S}_N)$).

9. The function $\mathrm{f}(\mathbf{S})$ is smooth and twice differentiable.

Let $\{\mathrm{f}(\mathbf{S}_N)\}_{N=1}^{\infty}$ be the sequence of PA estimator $\mathrm{f}(\mathbf{S}_N)$ (or any other estimator defined as a function of $S_{kN} \in \mathbf{S}_N$), where the regularly conditions 1 to 9 hold in addition to the following conditions:

(a) The sequence of estimators $\{\mathrm{f}(\mathbf{S}_N)\}_{N=1}^{\infty}$ is asymptotically unbiased for θ_N, that is

$$\lim_{N\to\infty} \mathbb{E}\left(\mathrm{f}(\mathbf{S}_N) - \mathrm{f}(\boldsymbol{\pi})\right) = 0. \tag{5.39}$$

This condition can be shown for any PA estimator $\mathrm{f}(\mathbf{S}_N)$ using the result (5.36).

(b) The variance of the sequence of estimators $\{\mathrm{f}(\mathbf{S}_N)\}_{N=1}^{\infty}$ goes to 0 as $N \to \infty$, that is

$$\lim_{N\to\infty} \mathbb{V}\left(\mathrm{f}(\mathbf{S}_N) \,|\, \mathcal{F}_N\right) = 0. \tag{5.40}$$

This condition is shown for any PA estimator $\mathrm{f}(\mathbf{S}_N)$ using the result (5.37), and it depends on the specific form of the PA estimator and sample design. See the following sections for specific forms of PA estimators.

Let $\left\{\hat{\mathbb{V}}\left(\mathrm{f}(\mathbf{S}_N)\right)\right\}_{N=1}^{\infty}$ be the sequence of variance estimators of a sequence of PA estimators $\left\{\mathrm{f}(\mathbf{S}_N)\right\}_{N=1}^{\infty}$ that meet the regularity conditions 1 to 9 in addition to conditions (a) and (b) and the following conditions:

(c) For each $\mathbf{y}_{N'} \in \{\mathcal{F}_N\}_{N=1}^{\infty}$ in the sequence of populations, $\mathcal{F}_{N'} \in \{\mathcal{F}_N\}_{N=1}^{\infty}$, the Euclidian norm of the Hadamard squared of $\mathbf{y}_{N'}$ is bounded, $\frac{1}{N}\left\|\mathbf{y}^{\odot 2}\right\|_2^2 = \mathcal{O}(1)$, as $N \to \infty$ where $\mathbf{y}^{\odot 2} = \mathbf{y} \odot \mathbf{y}$ (for consistency the variance estimator $\hat{\mathbb{V}}\left(\mathrm{f}(\mathbf{S}_N)\right)$).

(d) The sequence of estimators $\left\{\hat{\mathbb{V}}\left(\mathrm{f}(\mathbf{S}_N)\right)\right\}_{N=1}^{\infty}$ is asymptotically unbiased for $\left\{\mathbb{V}\left(\mathrm{f}(\mathbf{S}_N)\right)\right\}_{N=1}^{\infty}$, that is

$$\lim_{N\to\infty} \mathbb{E}\left(\hat{\mathbb{V}}\left(\mathrm{f}(\mathbf{S}_N)\right) - \mathbb{V}\left(\mathrm{f}(\mathbf{S}_N)\right)\right) = 0. \tag{5.41}$$

(e) The variance of the sequence of estimators $\left\{\hat{\mathbb{V}}\left(\mathrm{f}(\mathbf{S}_N)\right)\right\}_{N=1}^{\infty}$ goes to 0 as $N \to \infty$, that is

$$\lim_{N\to\infty} \mathbb{V}\left(\hat{\mathbb{V}}\left(\mathrm{f}(\mathbf{S}_N) \mid \mathcal{F}_N\right)\right) = 0. \tag{5.42}$$

Both conditions also depend on the specific form of the PA estimator and sample design. See the following sections for specific forms of PA estimators.

Assuming these regularity conditions hold, let $\{Z_N\}_{N=1}^{\infty}$ be the sequence of estimators defined as $Z_N = \mathrm{f}(\mathbf{S}_N) - \mathrm{f}(\boldsymbol{\pi}_N)$ where the function $\mathrm{f}: \mathbb{R}^N \to \mathbb{R}$ and

$$\mathbf{d}(\boldsymbol{\pi}) = \frac{\partial}{\partial \mathbf{S}} \mathrm{f}(\mathbf{S}) \Big|_{\mathbf{S}=\boldsymbol{\pi}},$$

is the vector of partial derivatives of $\mathrm{f}(\mathbf{S})$ evaluated at $\mathbf{S} = \boldsymbol{\pi}$. If $\mathbf{d}(\boldsymbol{\pi})$ is not equal to the zero vector and $\mathbf{d}(\boldsymbol{\pi})$ is continuous in the neighborhood of $\boldsymbol{\pi}$, and $(\mathbf{S}_N - \boldsymbol{\pi}_N) N_N^{-1/2} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Delta})$ (see Polansky 2011) then $Z_N \xrightarrow{d} \mathcal{N}\left(0, \left(\mathbf{d}(\boldsymbol{\pi})\right)^{\mathrm{T}} \boldsymbol{\Delta}\, \mathbf{d}(\boldsymbol{\pi})\right)$ as $N \to \infty$ (See Theorem 6.5 in Polansky 2011). As a result, the limiting distribution of the sequence of estimators $\left\{ \frac{Z_N}{\sqrt{\mathbb{V}(\mathrm{f}(\mathbf{S}))}} \right\}_{N=1}^{\infty}$ where $Z_N = \mathrm{f}(\mathbf{S}_N) - \mathrm{f}(\boldsymbol{\pi}_N)$ and where $\mathbb{V}(\mathrm{f}(\mathbf{S})) = \left(\mathbf{d}(\boldsymbol{\pi}_N)\right)^{\mathrm{T}} \boldsymbol{\Delta}_N\, \mathbf{d}(\boldsymbol{\pi}_N)$ is

$$\frac{\mathrm{f}(\mathbf{S}_N) - \mathrm{f}(\boldsymbol{\pi}_N)}{\sqrt{\mathbb{V}(\mathrm{f}(\mathbf{S}))}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $N \to \infty$. Using Slutsky's theorem, when $\mathbb{V}(\mathrm{f}(\mathbf{S})) = \left(\mathbf{d}(\boldsymbol{\pi}_N)\right)^{\mathrm{T}} \boldsymbol{\Delta}_N\, \mathbf{d}(\boldsymbol{\pi}_N)$ is estimated by $\hat{\mathbb{V}}(\mathrm{f}(\mathbf{S})) = \left(\hat{\mathbf{d}}(\boldsymbol{\pi}_N)\right)^{\mathrm{T}} \boldsymbol{\Delta}_N\, \hat{\mathbf{d}}(\boldsymbol{\pi}_N)$ then

$$\frac{\mathrm{f}(\mathbf{S}_N) - \mathrm{f}(\boldsymbol{\pi}_N)}{\sqrt{\hat{\mathbb{V}}(\mathrm{f}(\mathbf{S}))}} \xrightarrow{d} \mathcal{N}(0,1).$$

REMARK 5.3 The nonlinear PA estimators require solving more complex functions of $\mathbf{S}$ such as the inverse of link functions for GAMLSS models (e.g., exponential, negative inverse, and the inverse of the root square). The most complex expression is for nonlinear estimators with weights calibrated to the population and sample size, w_k for $k \in U$ which are also a function of $S_k \in \mathbf{S}$. Computing the TS approximations for these functions require derivatives of products of vectors/matrices using the matrix chain rule, the derivative of the inverse of matrices, and derivative of Hadamard products.

REMARK 5.4 Unlike estimating the parameter of nonlinear models that are solved iteratively (McCullagh & Nelder, 1989), the form of the PA estimator defied as

$$\hat{Y}_{PA} = \mathbf{w}^{\mathrm{T}} \left(\hat{\boldsymbol{\mu}}_{pa} \odot \mathbf{S} \right),$$

has always a closed form since $\hat{\boldsymbol{\mu}}_{pa} = \hat{\mathbb{E}}(\mathbf{y})$ where $\hat{\mathbb{E}}(\mathbf{y})$ depends on the density distribution of $\mathbf{y}$. Once the model parameters are estimated (they may be computed iteratively), they are plugged into the expression of $\boldsymbol{\mu} = \mathbb{E}(\mathbf{y})$ of the working model (see Section 1.5.1).

REMARK 5.5 The expressions (5.37) and (5.38) do not reflect the variability from the model selection. Modifications to these expressions to reflect the model selection variability will be the topic of future research.

REMARK 5.6 Some of the regularity conditions described above are identified based on the properties of the variance covariance matrix $\boldsymbol{\Delta}$ listed in Section 5.3, the redefinition of the sample design as a function of the discrete random variable for the sample membership indicator in Section 5.8 have not described before in the literature.

5.9.1 The Hájek Estimator

Let y be the variable of interest with a superpopulation model $\mathcal{M}_y$ where $y_k \sim \mathcal{N}\left(\beta,\sigma^2\right)$, $\beta \in \mathbb{R}_{\neq 0}$ is the location parameter. Let $\mathcal{F}$ be a finite population consisting of N *iid* realizations of $\mathcal{M}_y$. Let $\mathbf{S}$ be a random discrete vector that uniquely defines the sample design $p\left(\mathbf{S}=\mathbf{s}\right)$ with $\mathbb{E}\left(\mathbf{S}\right)=\boldsymbol{\pi}$ and $\mathbb{C}\left(\mathbf{S}\right)=\boldsymbol{\Delta}$ that meets the regularity conditions listed in Section 5.9 on page 252.

The PA estimator with this working model, the auxiliary variable 1, the total population N is the HJ estimator:

$$\hat{\bar{Y}}_{HJ} = \frac{\mathbf{d}^{\mathrm{T}}\left(\mathbf{y}\odot\mathbf{S}\right)}{\mathbf{d}^{\mathrm{T}}\mathbf{S}} = \hat{\beta}_{pmle} \tag{5.43}$$

The HJ estimator is a nonlinear function $\mathrm{f}:\mathbb{R}^N \to \mathbb{R}$ where $\mathrm{f}\left(\mathbf{S}\right)=\frac{a}{b}$, the numerator and denominator are linear functions of S_k with $a\left(\mathbf{S}\right)=\mathbf{d}^{\mathrm{T}}\left(\mathbf{y}\odot\mathbf{S}\right)$ and $b\left(\mathbf{S}\right)=\mathbf{d}^{\mathrm{T}}\mathbf{S}$. Using the results from Section 5.9, we approximate $\hat{\bar{Y}}_{HJ}$ by the first two terms of the TS of the function $\mathrm{f}_{\hat{\bar{Y}}_{HJ}}\left(\mathbf{S}\right)$ at the point $\mathbf{S}=\boldsymbol{\pi}$ as

$$\mathrm{f}_{\hat{\bar{Y}}_{HJ}}(\mathbf{S}) = \hat{\bar{Y}}_{HJ}\Big|_{\mathbf{S}=\boldsymbol{\pi}} + \left(\frac{\partial \hat{\bar{Y}}_{HJ}}{\partial \mathbf{S}}\right)^{\mathrm{T}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}} (\mathbf{S}-\boldsymbol{\pi}) + \mathcal{O}_p\left(\|\mathbf{S}-\boldsymbol{\pi}\|_2^2\right). \tag{5.44}$$

We focus on the term $\left.\frac{\partial \hat{\bar{Y}}_{HJ}}{\partial \mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}}$ which is a scalar-by-vector, partial, directional derivative with respect to the random vector $\mathbf{S}$. Using the chain rule for derivatives of matrices

$$\begin{aligned}
\left.\frac{\partial \hat{\bar{Y}}_{HJ}}{\partial \mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}} &= \left.\frac{\mathbf{d}\odot\mathbf{y}\odot\mathbf{1}}{\mathbf{d}^{\mathrm{T}}\mathbf{S}} - \frac{(\mathbf{d}\odot\mathbf{y})^{\mathrm{T}}\mathbf{S}}{\left(\mathbf{d}^{\mathrm{T}}\mathbf{S}\right)^2}(\mathbf{d}\odot\mathbf{1})\right|_{\mathbf{S}=\boldsymbol{\pi}} \\
&= \frac{\mathbf{d}\odot\mathbf{y}\odot\mathbf{1}}{N} - \frac{Y}{N^2}(\mathbf{d}\odot\mathbf{1}) = \frac{1}{N}\mathbf{d}\odot\left(\mathbf{y}-\mathbf{1}\bar{Y}\right), \\
&= \frac{1}{N}\mathbf{d}\odot\mathbf{e}
\end{aligned} \tag{5.45}$$

where $\mathbf{e} = \mathbf{y} - \bar{Y}$ is the vector of residuals of the model $\mathcal{M}_y$ fit to the entire population. The approximate variance of $\hat{\bar{Y}}_{HJ}$ is

$$\begin{aligned}
\mathbb{AV}\left(\hat{\bar{Y}}_{HJ}\right) &= \frac{1}{N^2}\,\mathbb{V}\left((\mathbf{d}\odot\mathbf{e})^{\mathrm{T}}(\mathbf{S}-\boldsymbol{\pi})\right) \\
&= \frac{1}{N^2}(\mathbf{d}\odot\mathbf{e})^{\mathrm{T}}\boldsymbol{\Delta}(\mathbf{d}\odot\mathbf{e})
\end{aligned} \tag{5.46}$$

The estimator of the variance $\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{HJ}\right)$, computed by replacing the unknown population quantities by their sample-based estimates, is

$$\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{HJ}\right) = \frac{1}{\hat{N}_{HT}^2}(\mathbf{d}\odot\breve{\mathbf{e}}\odot\mathbf{s})^{\mathrm{T}}\,\hat{\Delta}\,(\mathbf{d}\odot\breve{\mathbf{e}}\odot\mathbf{s}), \tag{5.47}$$

where $\breve{\mathbf{e}} = \left(\mathbf{y} - \hat{\bar{Y}}_{HJ}\right) \odot \mathbf{s}$ are the sample-based residuals of the PL model, $\hat{N}_{HT} = \mathbf{d}^{\mathrm{T}}\mathbf{s}$ and $\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$. The expression (5.47) matches the variance estimator of the HJ estimator in sampling books (Cochran, 1977).

Using the same arguments in Section 5.9, since the PA estimator $\hat{\bar{Y}}_{HJ}$ is a nonlinear function of $\mathbf{S}$ then the sequence of PA estimator $\left\{\hat{\bar{Y}}_{GREG,N}\right\}_{N=1}^{\infty}$ is design consistent of the population mean $\bar{Y}_N$. The limiting distribution of the sequence of estimators $\left\{\frac{\hat{\bar{Y}}_{HJ,N} - \hat{\bar{Y}}_N}{\sqrt{\mathbb{V}\left(\hat{\bar{Y}}_{HJ,N}\right)}}\right\}_{N=1}^{\infty}$ and $\left\{\frac{\hat{\bar{Y}}_{HJ,N} - \hat{\bar{Y}}_N}{\sqrt{\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{HJ,N}\right)}}\right\}_{N=1}^{\infty}$ is $\mathcal{N}(0,1)$.

5.9.2 The Classical Ratio Estimator

Let y be the variable of interest with a superpopulation model $\mathcal{M}_y$ with $y_k \sim \mathcal{N}\left(x_k\beta, x_k\sigma^2\right)$, where $x_k \in \mathbb{R}_{\neq 0}$ is the auxiliary variable, $\beta \in \mathbb{R}_{\neq 0}$ is the location parameter, and $X = \mathbf{1}^{\mathrm{T}} x \in \mathbb{R}$ is the population totals. Let $\mathcal{F}$ be a finite population consisting of N *iid* realizations of $\mathcal{M}_y$. Let $\mathbf{S}$ be a random discrete vector that uniquely defines the sample design $p(\mathbf{S} = \mathbf{s})$ with $\mathbb{E}(\mathbf{S}) = \boldsymbol{\pi}$ and $\mathbb{C}(\mathbf{S}) = \boldsymbol{\Delta}$ that meets the regularity conditions listed in Section 5.9 on page 252.

The PA estimator with the normal working model $\mathcal{M}_y$, the auxiliary variable x_k, and the population total X is the RA estimator:

$$\hat{Y}_{RA} = \frac{\mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{S})}{\mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{S})} = X\hat{\boldsymbol{\beta}}_{pmle}. \tag{5.48}$$

The RA estimator is a nonlinear function $\mathrm{f}: \mathbb{R}^N \to \mathbb{R}$ where $\mathrm{f}(\mathbf{S}) = \frac{a}{b}$, the numerator and denominator are linear functions of S_k with $a(\mathbf{S}) = \mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{S})$ and $b(\mathbf{S}) = \mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{S})$. Using the results from Section in Section 5.9, we approximate $\hat{\hat{Y}}_{RA}$ by the first two terms of the TS of the function $\mathrm{f}_{\hat{Y}_{RA}}(\mathbf{S})$ at the point $\mathbf{S} = \boldsymbol{\pi}$ as

$$\mathrm{f}_{\hat{Y}_{RA}}(\mathbf{S}) = \hat{Y}_{RA}\Big|_{\mathbf{S}=\boldsymbol{\pi}} + \left(\frac{\partial \hat{Y}_{RA}}{\partial \mathbf{S}}\right)^{\mathrm{T}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}} (\mathbf{S} - \boldsymbol{\pi}) + \mathcal{O}_p\left(\|\mathbf{S} - \boldsymbol{\pi}\|_2^2\right). \tag{5.49}$$

We focus on the term $\left.\frac{\partial \hat{Y}_{RA}}{\partial \mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}}$ which is a scalar-by-vector, partial, directional derivative with respect to the random vector $\mathbf{S}$. Using the chain rule for derivatives of matrices

$$\begin{aligned}
\left.\frac{\partial \hat{Y}_{RA}}{\partial \mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}} &= \left.\left(\frac{(\mathbf{d} \odot \mathbf{y})^{\mathrm{T}}}{\mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{S})} - \frac{\mathbf{d}^{\mathrm{T}}(\mathbf{y} \odot \mathbf{S})(\mathbf{d} \odot \mathbf{x})^{\mathrm{T}}}{\left(\mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{S})\right)^2}\right)\right|_{\mathbf{S}=\boldsymbol{\pi}} \\
&= \frac{(\mathbf{d} \odot \mathbf{y})^{\mathrm{T}}}{X} - \frac{Y(\mathbf{d} \odot \mathbf{x})^{\mathrm{T}}}{X^2} \\
&= \frac{1}{X}\mathbf{d} \odot \left(\mathbf{y} - \mathbf{x}\frac{Y}{X}\right) = \frac{1}{X}\mathbf{d} \odot (\mathbf{y} - \mathbf{x}R) \\
&= \frac{1}{X}\mathbf{d} \odot \mathbf{e}
\end{aligned} \tag{5.50}$$

where $\mathbf{e} = \mathbf{y} - \bar{Y}$ is the vector of residuals of the model $\mathcal{M}_y$ fit to the complete population. The approximate variance of $\hat{Y}_{RA}$ is

$$\mathbb{AV}\left(\hat{Y}_{RA}\right) = \frac{1}{X^2}\left(\mathbf{d} \odot \mathbf{e}\right)^{\mathrm{T}} \boldsymbol{\Delta}\left(\mathbf{d} \odot \mathbf{e}\right) \tag{5.51}$$

The estimator of the variance $\hat{\mathbb{V}}\left(\hat{Y}_{RA}\right)$, computed by replacing the unknown population quantities by their sample-based estimates, is

$$\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{RA}\right) = \frac{1}{\hat{X}_{HT}^2}\left(\mathbf{d} \odot \breve{\mathbf{e}} \odot \mathbf{s}\right)^{\mathrm{T}} \hat{\boldsymbol{\Delta}}\left(\mathbf{d} \odot \breve{\mathbf{e}} \odot \mathbf{s}\right), \tag{5.52}$$

where $\breve{\mathbf{e}} = \left(\mathbf{y} - \mathbf{x}\hat{R}_{HT}\right) \odot \mathbf{s}$ are the sample-based residuals of the PL model, $\hat{R}_{HT} = \dfrac{\mathbf{d}^{\mathrm{T}}\left(\mathbf{y} \odot \mathbf{s}\right)}{\mathbf{d}^{\mathrm{T}}\left(\mathbf{x} \odot \mathbf{s}\right)}$, $\hat{X}_{HT} = \mathbf{d}^{\mathrm{T}}\left(\mathbf{x} \odot \mathbf{s}\right)$, and $\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$. The expression (5.52) matches the variance estimator of the RA estimator in sampling books (Cochran, 1977).

Using the same arguments in Section 5.9, since the PA estimator $\hat{\bar{Y}}_{RA}$ is a nonlinear function of $\mathbf{S}$ then the sequence of PA estimator $\left\{\hat{\bar{Y}}_{RA,N}\right\}_{N=1}^{\infty}$ is design consistent of the population mean $\bar{Y}_N$. The limiting distribution of the sequence of estimators $\left\{\dfrac{\hat{\bar{Y}}_{RA} - \hat{\bar{Y}}_N}{\sqrt{\mathbb{V}\left(\hat{\bar{Y}}_{RA}\right)}}\right\}_{N=1}^{\infty}$ and $\left\{\dfrac{\hat{\bar{Y}}_{RA} - \hat{\bar{Y}}_N}{\sqrt{\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{RA}\right)}}\right\}_{N=1}^{\infty}$ is $\mathcal{N}(0,1)$.

5.9.3 The Linear PA Estimator (GREG)

Let y be the variable of interest with a superpopulation model $\mathcal{M}_y$ where $y_k \sim \mathcal{N}\left(\mathbf{x}_k\boldsymbol{\beta},\sigma^2\right)$, $\mathbf{x}_k = \left(x_{k1},\ldots,x_{kP}\right) \in \mathbb{R}^{1\times P}$ is the vector of auxiliary variables, $\boldsymbol{\beta} = \left(\beta_1,\ldots,\beta_P\right)^{\mathrm{T}} \in \mathbb{R}^{P\times 1}$ is the vector with the location parameters, and $\mathbf{X} = \mathbf{1}^{\mathrm{T}}\mathbf{x} \in \mathbb{R}^{1\times P}$ is the vector of the population totals of the auxiliary variables $\mathbf{x}$. Let $\mathcal{F}$ be a finite population consisting of N *iid* realizations of $\mathcal{M}_y$. Let $\mathbf{S}$ be a random discrete vector that uniquely defines the sample design $p\left(\mathbf{S}=\mathbf{s}\right)$ with $\mathbb{E}\left(\mathbf{S}\right)=\boldsymbol{\pi}$ and $\mathbb{C}\left(\mathbf{S}\right)=\boldsymbol{\Delta}$ that meets the regularity conditions listed in Section 5.9 on page 252.

The PA estimator of the population mean $\bar{Y}$ based on the model $\mathcal{M}_y$ is

$$\hat{\bar{Y}}_{GREG} = \frac{1}{N}\mathbf{X}\hat{\boldsymbol{\beta}}_{pmle} = \frac{1}{N}\mathbf{X}\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}}, \tag{5.53}$$

where $\hat{\mathbf{T}}_{\mathbf{xx}} = \left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{x} \in \mathbb{R}^{P\times P}$, $\hat{\mathbf{T}}_{\mathbf{xy}} = \left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{y} \in \mathbb{R}^{P\times 1}$, and $\hat{\boldsymbol{\beta}}_{pml} = \hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}} \in \mathbb{R}^{P\times 1}$. This expression 5.53) matches the GREG estimator in Särndal, Swensson, & Wretman (1992).

The variance of $\hat{\bar{Y}}_{GREG}$ is

$$\mathbb{V}\left(\hat{\bar{Y}}_{GREG}\right) = \frac{1}{N^2}\mathbf{X}^{\mathrm{T}}\mathbb{C}\left(\hat{\boldsymbol{\beta}}_{pmle}\right)\mathbf{X}.$$

Using the results from Section A.3.2 for the variance-covariance $\mathbb{C}\left(\hat{\boldsymbol{\beta}}_{pmle}\right)$ in (A.22), the approximate variance of $\hat{\bar{Y}}_{GREG}$ is

$$\mathbb{AV}\left(\hat{\bar{Y}}_{GREG}\right) = \frac{1}{N^2}\mathbf{X}^{\mathrm{T}}\mathbf{T}_{\mathbf{xx}}^{-1}\left(\mathbf{x}\odot\mathbf{d}\odot\mathbf{e}\right)^{\mathrm{T}}\Delta\left(\mathbf{x}\odot\mathbf{d}\odot\mathbf{e}\right)\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{X}. \tag{5.54}$$

The expression of $\mathbb{AV}\left(\hat{\bar{Y}}_{GREG}\right)$ matches those in Särndal, Swensson, & Wretman (1989), Binder (1996), and Demnati & Rao (2004) which includes the g-weights. This expression does not reflect the effect of the model selection on the variance estimator.

The variance estimator $\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{GREG}\right)$, computed by replacing the unknown population quantities by their sample-based estimates, is

$$\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{GREG}\right) = \frac{1}{N^2}\mathbf{X}\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\left(\mathbf{x}\odot\mathbf{d}\odot\breve{\mathbf{e}}\odot\mathbf{s}\right)^{\mathrm{T}}\hat{\Delta}\left(\mathbf{x}\odot\mathbf{d}\odot\breve{\mathbf{e}}\odot\mathbf{s}\right)\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\mathbf{X}^{\mathrm{T}}, \tag{5.55}$$

where $\breve{\mathbf{e}} = \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\odot\mathbf{s}$ are the sample-based residuals of the PL model, $\hat{\mathbf{T}}_{\mathbf{xx}}$ is the matrix of the HT estimates of the cross product $\mathbf{x}^{\mathrm{T}}\mathbf{x}$, and $\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}\oslash\boldsymbol{\Pi}$. Using the same arguments in Section 5.9, since the PA estimator $\hat{\bar{Y}}_{GREG}$ is a nonlinear function of $\mathbf{S}$ then the sequence of PA estimator $\left\{\hat{\bar{Y}}_{GREG,N}\right\}_{N=1}^{\infty}$ is design consistent of the population mean $\bar{Y}_N$. The limiting distribution of the sequence of estimators

$$\left\{\frac{\hat{\bar{Y}}_{GREG,N} - \hat{\bar{Y}}_N}{\sqrt{\mathbb{V}\left(\hat{\bar{Y}}_{GREG,N}\right)}}\right\}_{N=1}^{\infty} \text{ and } \left\{\frac{\hat{\bar{Y}}_{GREG,N} - \hat{\bar{Y}}_N}{\sqrt{\hat{\mathbb{V}}\left(\hat{\bar{Y}}_{GREG,N}\right)}}\right\}_{N=1}^{\infty} \text{ is } \mathcal{N}(0,1).$$

5.9.4 The Nonlinear PA Estimator for Poisson Model with the log Link Function

Let y the variable of interest with a superpopulation model $\mathcal{M}_y$ where $y_k \sim Poisson(\lambda)$, $\mathbb{E}(y_k) = \lambda$, $\log(\lambda) = \mathbf{x}_k\boldsymbol{\beta}$, $\mathbf{x}_k = (x_{k1},\ldots,x_{kP}) \in \mathbb{R}^{1\times P}$ is the vector of auxiliary variables, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_P)^{\mathrm{T}} \in \mathbb{R}^{P\times 1}$ is the vector with the location parameters, and $\mathbf{X} = \mathbf{1}^{\mathrm{T}}\mathbf{x} \in \mathbb{R}^{1\times P}$ is the vector of the population totals of the auxiliary variables $\mathbf{x}$. Let $\mathcal{F}$ be a finite population consisting of N *iid* realizations of $\mathcal{M}_y$. Let $\mathbf{S}$ be a random discrete vector that uniquely defines the sample design $p(\mathbf{S} = \mathbf{s})$ with $\mathbb{E}(\mathbf{S}) = \boldsymbol{\pi}$ and $\mathbb{C}(\mathbf{S}) = \boldsymbol{\Delta}$ that meets the regularity conditions listed in Section 5.9 on page 252.

The PA estimator of the total Y based on $\mathcal{M}_y$ with Poisson model, the location parameter $\theta_\beta = \mathbf{x}\boldsymbol{\beta}$, log link function, the auxiliary variables $\mathbf{x}$, and population totals $\mathbf{X}$, is

$$\hat{Y}_{PO} = \mathbf{d}^{\mathrm{T}}\left(\hat{\boldsymbol{\mu}}_{pa} \odot \mathbf{S}\right), \tag{5.56}$$

where $\hat{\boldsymbol{\mu}}_{pa}$ is the vector PA adjusted fitted mean of the model where $\hat{\mu}_{pa} = \exp\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{PA} \odot \mathbf{S}\right)$ and $\mathbf{d} = [d_k] = \boldsymbol{\pi}^{\odot -1} = \left[\frac{1}{\pi_k}\right] \in \mathbb{R}^{N\times 1}$ are the sampling weights.

Using the results from Section 5.9, we approximate $\hat{Y}_{PO}$ by the first two terms of the TS of the function $\mathrm{f}_{\hat{Y}_{PO}}(\mathbf{S})$ at the point $\mathbf{S}=\boldsymbol{\pi}$ as

$$\mathrm{f}_{\hat{Y}_{PO}}(\mathbf{S}) = \hat{Y}_{PO}\Big|_{\mathbf{S}=\boldsymbol{\pi}} + \left(\frac{\partial \hat{Y}_{PO}}{\partial \mathbf{S}}\right)^{\mathrm{T}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}} (\mathbf{S}-\boldsymbol{\pi}) + \mathcal{O}_p\left(\|\mathbf{S}-\boldsymbol{\pi}\|_2^2\right). \tag{5.57}$$

We focus on the term $\left.\frac{\partial \hat{Y}_{PO}}{\partial \mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}}$ which is a scalar-by-vector, partial, directional derivative with respect to the random vector $\mathbf{S}$. To compute the approximate variance, we use (5.38) as

$$\begin{aligned}\mathbb{AV}\left(\hat{Y}_{PO}\right) = &\left(\mathbf{d}\odot\hat{\boldsymbol{\mu}}_{mle}\right)^{\mathrm{T}}\boldsymbol{\Delta}\left(\mathbf{d}\odot\hat{\boldsymbol{\mu}}_{mle}\right)\\ &+\left(\frac{\partial\hat{\boldsymbol{\mu}}_{pa}(\boldsymbol{\pi})}{\partial\mathbf{S}}\right)^{\mathrm{T}}\boldsymbol{\Delta}\left(\frac{\partial\hat{\boldsymbol{\mu}}_{pa}(\boldsymbol{\pi})}{\partial\mathbf{S}}\right).\\ &+2\left(\mathbf{d}\odot\hat{\boldsymbol{\mu}}_{mle}\right)^{\mathrm{T}}\boldsymbol{\Delta}\left(\frac{\partial\hat{\boldsymbol{\mu}}_{pa}(\boldsymbol{\pi})}{\partial\mathbf{S}}\right)\end{aligned} \tag{5.58}$$

When computing the variance, we distinguish the following terms

- V_1 is the component of the variance of the HT estimator with the variable $\hat{\boldsymbol{\mu}}_{pa}$,

$$V_1 = \left(\mathbf{d}\odot\hat{\boldsymbol{\mu}}_{mle}\right)^{\mathrm{T}}\boldsymbol{\Delta}\left(\mathbf{d}\odot\hat{\boldsymbol{\mu}}_{mle}\right). \tag{5.59}$$

- V_2 is the component of variance for the linearized part of $\hat{\boldsymbol{\mu}}_{pa}$ represented by $\frac{\partial\hat{\boldsymbol{\mu}}_{pa}}{\partial\mathbf{S}}$ as

$$V_2 = \left(\frac{\partial\hat{\boldsymbol{\mu}}_{pa}}{\partial\mathbf{S}}\right)^{\mathrm{T}}\boldsymbol{\Delta}\frac{\partial\hat{\boldsymbol{\mu}}_{pa}}{\partial\mathbf{S}}. \tag{5.60}$$

Since $\frac{\partial \hat{\boldsymbol{\mu}}_{pa,k}}{\partial \mathbf{S}} = \hat{\boldsymbol{\mu}}_{pa,k} \frac{\partial \mathbf{x}_k \hat{\boldsymbol{\beta}}_{pa}}{\partial \mathbf{S}}$ and using the results from Section A.3.3, this component can be decomposed in the following components.

- V_{21} is the component of variance from the model fit $\mathrm{g}^{-1}\left(\hat{\boldsymbol{\mu}}_{mle}\right) = \mathbf{x}\hat{\boldsymbol{\beta}}_{mle}$ with the residuals $\mathbf{e} = \hat{\boldsymbol{\mu}}_{mle} - \mathrm{g}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{mle}\right)$ as

$$V_{21} = \left(\hat{\boldsymbol{\mu}}_{mle} \odot \mathbf{x} \odot \mathbf{e}\right)^{\mathrm{T}} \boldsymbol{\Delta}\left(\hat{\boldsymbol{\mu}}_{mle} \odot \mathbf{x} \odot \mathbf{e}\right). \tag{5.61}$$

- V_{22} is the component of variance from PA adjustment $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$ made to the regression coefficients $\hat{\boldsymbol{\beta}}_{pmle}$ as

$$V_{22} = \left(\hat{\boldsymbol{\mu}}_{mle} \odot \frac{\hat{\beta}_{mle,p}}{X_p} \mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e}\right)^{\mathrm{T}} \boldsymbol{\Delta}\left(\hat{\boldsymbol{\mu}}_{mle} \odot \frac{\hat{\beta}_{mle,p}}{X_p} \mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e}\right), \tag{5.62}$$

or $V_{22} = 0$ if $p \neq q \in \{1,...,P\}$.

- V_{23} is the component of variance from the correlation between the PA adjustment and the model fit $\mathrm{g}^{-1}\left(\hat{\mu}_{mle}\right) = \mathbf{x}\hat{\boldsymbol{\beta}}_{mle}$ as

$$\begin{aligned} V_{23} = &\left(\hat{\boldsymbol{\mu}}_{mle} \odot \frac{\hat{\beta}_{mle,p}}{X_p} \mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e}\right)^{\mathrm{T}} \boldsymbol{\Delta}\left(\hat{\boldsymbol{\mu}}_{mle} \odot \mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e}\right) \\ &+\left(\hat{\boldsymbol{\mu}}_{mle} \odot \mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e}\right)^{\mathrm{T}} \boldsymbol{\Delta}\left(\hat{\boldsymbol{\mu}}_{mle} \odot \frac{\hat{\beta}_{mle,p}}{X_p} \mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e}\right) \end{aligned}, \tag{5.63}$$

or $V_{23} = 0$ if $p \neq q \in \{1,...,P\}$.

- V_3 is the component of variance form variance-covariance between the HT estimator with the variable $\hat{\boldsymbol{\mu}}_{pa}$, the PA adjustment, and the model fit $\mathrm{g}^{-1}\left(\hat{\mu}_{mle}\right)=\mathbf{x}\hat{\boldsymbol{\beta}}_{mle}$ as

$$\begin{aligned} V_3 = & 2\left(\hat{\boldsymbol{\mu}}_{mle}\odot\mathbf{x}\odot\mathbf{e}\right)^{\mathrm{T}}\boldsymbol{\Delta}\left(\hat{\boldsymbol{\mu}}_{mle}\odot\frac{\hat{\beta}_{mle,p}}{X_p}\mathbf{x}_p\odot\mathbf{d}\odot\mathbf{e}\right) \\ & +2\left(\hat{\boldsymbol{\mu}}_{mle}\odot\mathbf{x}\odot\mathbf{e}\right)^{\mathrm{T}}\boldsymbol{\Delta}\left(\hat{\boldsymbol{\mu}}_{mle}\odot\mathbf{x}_p\odot\mathbf{d}\odot\mathbf{e}\right) \\ & +2\left(\hat{\boldsymbol{\mu}}_{mle}\odot\frac{\hat{\beta}_{mle,p}}{X_p}\mathbf{x}_p\odot\mathbf{d}\odot\mathbf{e}\right)^{\mathrm{T}}\boldsymbol{\Delta}\left(\hat{\boldsymbol{\mu}}_{mle}\odot\mathbf{x}_p\odot\mathbf{d}\odot\mathbf{e}\right) \end{aligned}, \qquad (5.64)$$

or $V_3=0$ if $p\neq q\in\{1,...,P\}$.

The approximate variance is the sum of all these components as

$$\mathbb{AV}\left(\hat{Y}_{PO}\right)=V_1+V_{21}+V_{22}+V_{23}+V_3. \qquad (5.65)$$

The variance estimator $\hat{\mathbb{V}}\left(\hat{Y}_{PO}\right)$ is computed by replacing the unknown population quantities by their sample-based estimates, that is $\breve{\mathbf{e}}=\left(\mathbf{y}-\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\odot\mathbf{s}$, $\hat{\boldsymbol{\mu}}_{mle}$ by $\hat{\boldsymbol{\mu}}_{pmle}$, $\mathbf{T_{xx}}$ by $\hat{\mathbf{T}}_{\mathbf{xx}}$, and $\hat{\boldsymbol{\Delta}}=\boldsymbol{\Delta}\oslash\boldsymbol{\Pi}$.

The variance estimator $\hat{\mathbb{C}}\left(\hat{\beta}_{pa,p},\hat{\beta}_{pa,q}\right)$ is computed by replacing the unknown population quantities by their sample-based estimates, that is, $\mathbf{e}=\hat{\boldsymbol{\mu}}_{mle}-\mathrm{g}^{-1}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{mle}\right)$ by $\breve{\mathbf{e}}=\left(\hat{\boldsymbol{\mu}}_{pmle}-\mathrm{g}^{-1}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)\odot\mathbf{s}$, $\mathbf{T_{xx}}$, by $\hat{\mathbf{T}}_{\mathbf{xx}}$, the matrix of the HT estimates of the population of the cross product totals of $\mathbf{x}$, and $\boldsymbol{\Delta}$ by $\hat{\boldsymbol{\Delta}}=\boldsymbol{\Delta}\oslash\boldsymbol{\Pi}$.

5.10 Defining a Sequence for the Population y in Survey Sampling Asymptotic Theory

In this section, we elaborate on some conditions for design consistency that are not often discussed in the current literature. In standard statistical asymptotic theory, the large sample properties of estimators and statistical tests are assessed assuming that sample size n goes to infinity (Polansky, 2011). The standard approach for the study of the asymptotic properties in surveys was established in Isaki & Fuller (1982), and numerous papers use this approach. Isaki & Fuller's setup assumes an indexed sequence of nested finite populations $\{\mathcal{F}_N\}_{N=1}^{\infty}$ with labels $\{U_N = \{1,...,N_N\}\}_{N=1}^{\infty}$ and associated probability samples $\{A_N\}_{N=1}^{\infty}$ drawn according to a sample design $\{p_N(A_N = a_N)\}_{N=1}^{\infty}$ from each finite population in the sequence. In this setting, both the finite population size N_N and sample size n_N increase to infinity but the ratio is finite, since by definition, $\lim_{\substack{N\to\infty \\ n\to\infty}} \frac{n_N}{N_N} = f_N = f \in (0,1)$ with conditions such as $\limsup_{N\to\infty} \frac{1}{N_N} \sum_{k\in U_N} y_k^2 < \infty$ and $\limsup_{N\to\infty} \left\{ n \max_{k\neq l\in U_N} \left|\Delta_{N,kl}\right| \right\} < \infty$ (for the variance of the HT estimator to converge to 0), or $\min_{k,l\in U_N} \{\pi_{Nkl}\} \geq \lambda > 0$ and $\limsup_{N\to\infty} \frac{1}{N_N} \sum_{k\in U_N} y_k^4 < \infty$ (for the variance estimate of the HT estimator).

Although this approach is sound, the consistency of the sequence of estimators depends on the sequence $\{\mathbf{y}_N\}_{N=1}^{\infty}$ and $\{\boldsymbol{\pi}_N\}_{N=1}^{\infty}$ which are not explicitly defined, except for $\mathbf{y}_N$ which is assumed to have finite population moments.

A complete study of the asymptotic properties of an estimator requires examining the limiting behavior of quantities that are used to compute the estimator. For example, consider the expected value of the estimator described in Section 5.7,

$$\mathrm{Z}(\mathbf{S}) = \mathbb{E}\left(\frac{1}{N}\mathbf{a}^{\mathrm{T}}\mathbf{S}\right) = \frac{1}{N}\mathbf{a}^{\mathrm{T}}\boldsymbol{\pi},$$

with the corresponding sequence of estimators $\{Z_N\}_{N=1}^{\infty}$. In order to determine the large sample properties of Z, we need to define the limiting behavior of $\mathbf{a}$, $\boldsymbol{\pi}$, and $\boldsymbol{\Delta}$, as $N \to \infty$. When N increases, the size of the vector $\mathbf{S}_N$ also increases. The increasing size of $\mathbf{S}_N$ affects $\boldsymbol{\pi}_N$, $\boldsymbol{\Delta}_N$, and $\mathbf{a}_N$; they also increase in size. For example, the condition that $\boldsymbol{\pi}_N$ is finite leads to $n_N = \mathbf{1}_N^T \boldsymbol{\pi}_N < \infty$, where n_N. the sample size in the population N_N, is not sufficient since it does not describe the relationship between n_N and N_N as N_N increases.

A way to solve this dependency is by linking the limiting behavior of n_N and N_N as $\lim_{\substack{N\to\infty \\ n\to\infty}} \frac{n_N}{N_N} = f \in (0,1)$. This limiting sampling rate also implicitly links the behavior of $\boldsymbol{\pi}_N$ and N_N. The sum of $\boldsymbol{\pi}$ is the same order of N, that is, the term $\frac{\mathbf{1}_N^{\mathrm{T}}\boldsymbol{\pi}_N}{N_N}$ is

$\mathcal{O}(1)$. This order means the sum of the elements of $\boldsymbol{\pi}_N$ can go to infinity, but it must be of the same order of N. This order also implies that the sample size n cannot be set separately since it depends on the design. When we indicate that $N \to \infty$ and $n \to \infty$ such as $n/N \to f$, what we mean is that $\mathbf{1}_N^{\mathrm{T}} \boldsymbol{\pi}_N \to \infty$, so the proprieties of $\mathbf{S}$ are being defined since by definition, $\min_{k \in U} \arg\{\pi_k\} > 0$ and $\boldsymbol{\Delta}$ must meet the properties of Hermitian matrices in addition to the properties of the type sample design (See Sections 5.2 and 5.3). Both properties also imply that $\pi_{N,\max k} = \max_{k \in N_N} \arg(\boldsymbol{\pi}_N) = \mathcal{O}(N)$. However, the limiting behavior of Z_N also depends on $\mathbf{a}_N$, which may not be related to $\boldsymbol{\pi}_N$. In other to keep the order $\mathcal{O}(1)$ in $\frac{\mathbf{y}_N^{\mathrm{T}} \boldsymbol{\pi}_N}{N_N}$, the sum $\mathbf{y}_N^{\mathrm{T}} \boldsymbol{\pi}_N$ needs to be $\mathcal{O}(N)$. This order is achieved when $\mathbf{y}_N^{\mathrm{T}} \mathbf{1}_N = \mathcal{O}(1)$, since $\boldsymbol{\pi}_N$ is $\mathcal{O}(N)$.

Define $\{\mathbf{y}_N\}_{N=1}^{\infty}$ as a sequence of real constants where $\bar{Y}_N = \mathcal{O}\left(N^P\right)$, that is that the mean of the population increases in $\{\mathcal{F}_N\}_{N=1}^{\infty}$ but it is bounded by $\mathcal{O}\left(N^p\right)$. Let $\left\{\hat{\bar{Y}}_{HT,N}\right\}_{N=1}^{\infty}$ be a sequence of HT estimators of $\bar{Y}_N$ from samples drawn according to the sample designs $\left\{p_N\left(A_N = a_N\right)\right\}_{N=1}^{\infty}$ from the populations $\{\mathcal{F}_N\}_{N=1}^{\infty}$. Let

$\left\{\mathbb{V}\left(\hat{\bar{Y}}_{HT,N} \mid \mathcal{F}_N\right)\right\}_{N=1}^{\infty}$ be the sequence of variances of $\hat{\bar{Y}}_{HT,N}$. From Section 5.8.1, the upper bound of $\mathbb{V}\left(\bar{Y}_{N,HT} \mid \mathcal{F}_N\right)$ is $\mathbb{V}\left(\bar{Y}_{N,HT} \mid \mathcal{F}_N\right) \leq K_N \bar{Y}_N^2$, since $\frac{1}{N}\|\mathbf{y}_N\|_2^2 \leq N\bar{Y}_N^2$.

The value of p such as $\mathbb{V}\left(\bar{Y}_{N,HT} \mid \mathcal{F}_N\right)$ does not converge, e.g., $\mathbb{V}\left(\bar{Y}_{N,HT} \mid \mathcal{F}_N\right) \geq \mathcal{O}(1)$, is obtained by solving the expression $\left(\mathcal{O}\left(N^p\right)\right)^2 \geq \mathcal{O}(1)$. If $\bar{Y}_N$ grows at the same rate as the population size N, e.g., $p = 0$, then $\mathbb{V}\left(\bar{Y}_{N,HT} \mid \mathcal{F}_N\right)$ does not converge. If $-\frac{1}{2} < p < 1$ then $\mathbb{V}\left(\bar{Y}_{N,HT} \mid \mathcal{F}_N\right)$ converges at a slower rate than $\mathcal{O}\left(N^{-1}\right)$ and if $p < -\frac{1}{2}$, it converges at a faster rate than $\mathcal{O}_p\left(N^{-1}\right)$. If the mean of the population stays constant as the population increases, then $\bar{Y}_N = \mathcal{O}(\log N) \leq \mathcal{O}\left(N^{-1}\right)$, then $\mathbb{V}\left(\bar{Y}_{HT,N}\right)$ converges at a much faster rate to zero than $\mathcal{O}\left(N^{-1}\right)$. One implicit assumption in this development is $K_N = \mathcal{O}(1)$ as $N \to \infty$.

These results provide guidelines for the study of the asymptotic properties of the estimators through simulations, since they describe how the different finite populations can be generated depending on the relationship between the N, $\boldsymbol{\pi}$, and $\mathbf{y}$ as the population size increases. Notice that we assume that the model does not change as $N \to \infty$. See McConville, Breidt, Lee, & Moisen (2017) for the case that the number of regression coefficients increases as $N \to \infty$.

Chapter 6 Final Comments

In this paper, we introduce the PA framework for estimation with full response. The PA framework is a methodology for producing efficient estimators by targeting the auxiliary variables related to the outcome or outcomes. A key application is variable selection for efficient calibration estimators.

Despite using models, the PA estimators are model-assisted (in contrast to model-dependent), asymptotically consistent, and their properties do not depend on whether the model holds or not. Inferences depend on the sampling strategy or sample design used to draw the sample.

All PA estimators are sums of expanded estimated adjusted means of models where the model parameters for location, scale, and shape are functions of linear regressions of the auxiliary variables. Different auxiliary variables and model parameters produce different PA estimators. The PA framework establishes a link between standard statistical theory and design-based estimation. The approach justifies the use of standard statistical modeling for building working models and estimators within the design-based paradigm. The modeling approaches provide a metric for identifying the functional form of the model and for selecting the relevant auxiliary variables of the model. Current model-assisted approaches do not provide such metrics.

The PA estimators are derived algorithmically from the observed sample. Since the PA algorithm evaluates a pool of models, it avoids reliance on specifying a single working model with a specific set of auxiliary variables without a clear rationale. Since the metric and model are well defined, the creation of algorithmic PA estimators can be fully automated. Current practice does not provide such tools.

If the working model and set of auxiliary variables are specified, then the PA methodology reproduces most classical survey estimators using the algebraic PA approach.

Even complex estimators such as the Deville's Euclidian distance calibrated estimator and Särndal's generalized regression estimator (GREG) are also special cases of PA estimators. Furthermore, as illustrated in examples, new design-based estimators can be derived or engineered when the working model and auxiliary variables are specified.

The focus of the PA framework presented here is the estimation with full response, but the proposed methodology is a stepping-stone towards the development of estimation in the presence of nonresponse.

The presented framework also can be extended to estimators for domains, estimators from a cluster and two-stage designs, and estimators for other population characteristics such as the population distribution function and order statistics (i.e., quantiles and median).

The loss function in the current implementation of the PA algorithm is based on a sample-based version of the AIC, although other metrics for goodness of fit could be used.

A very important line of research is accounting for model selection. The challenge is to ensure statistical inference is valid following PA variable selection.

Finally, the approach we have used treats the sample design as having a multinomial distribution, and design-based estimators are functions of the random vector of the membership indicators. This approach provides a different way to study the survey sampling estimation theory. By using matrix notation and matrix operations, the PA framework facilitates obtaining asymptotic properties by relying on results from standard statistical theory.

We believe this approach is better suited for concepts such as the asymptotic relative efficiency of design-based estimators, providing insights on the efficiency of estimators when the sample sizes are small.

Appendix A Supplemental Plots and Proofs

A.1 Figures for Simulation Study in Section 2.2

This section contains the plots with relative bias (RB) and relative efficiency (RB) of the scenarios in the simulation study described in Section 2.3 on the evaluation of the performance of linear and nonlinear algorithmic PA estimators (see Section A.4 for the definitions of empirical measures). There are nine figures grouped by the distribution of the population:

Population	Figures
Binomial (binary data)	A.1 to A.3
Poisson (count data)	A.4 to A.6
Gamma: Continuous positive data with a constant coefficient of variation	A.7 to A.9

Each figure shows the RB and RE of estimators of the total population under repeated sampling (100,000 draws) from sample sizes drawn with a constant sampling rate ranging from 100 to 1000 cases with a fixed sampling rate of 0.05. In each plot, the vertical axis corresponds to the sample size from 100 to 1,000. The vertical axis on the left plot is RB while on the right is the RE; both are shown in percentage points. In each figure, the rows show the estimates by model strength measured by $\rho_{\eta X}$. The top plots correspond to low ($\rho_{\eta X} = 0.3$), the middle plots are medium ($\rho_{\eta X} = 0.6$), and the bottom plots are high ($\rho_{\eta X} = 0.9$). Within each population, the first figure shows the results for samples drawn

using simple random sampling (SRS), the second for sampling with probability proportional to size (PPS) and the last for Poisson sampling (PO).

Additional information on the factors and models for this study is found in Tables 2.3 and 2.5. The expressions of the estimators are listed in Table 2.4. The following symbols identify the estimators on plots A-1 to A-9:

Estimator	Symbol
Hájek	HJ
Model Calibrated	MC
Generalized Regression	GREG
Algorithmic Linear Parametric	LNPA
Algorithmic Nonlinear Parametric	NLPA
Algorithmic Non-linear calibrated PA	NLCA

Figure A.1 Relative Bias (RB) and Relative efficiency (RE) of seven estimators as a function of the sample size for the population with a Bernoulli distribution with SRS designs.

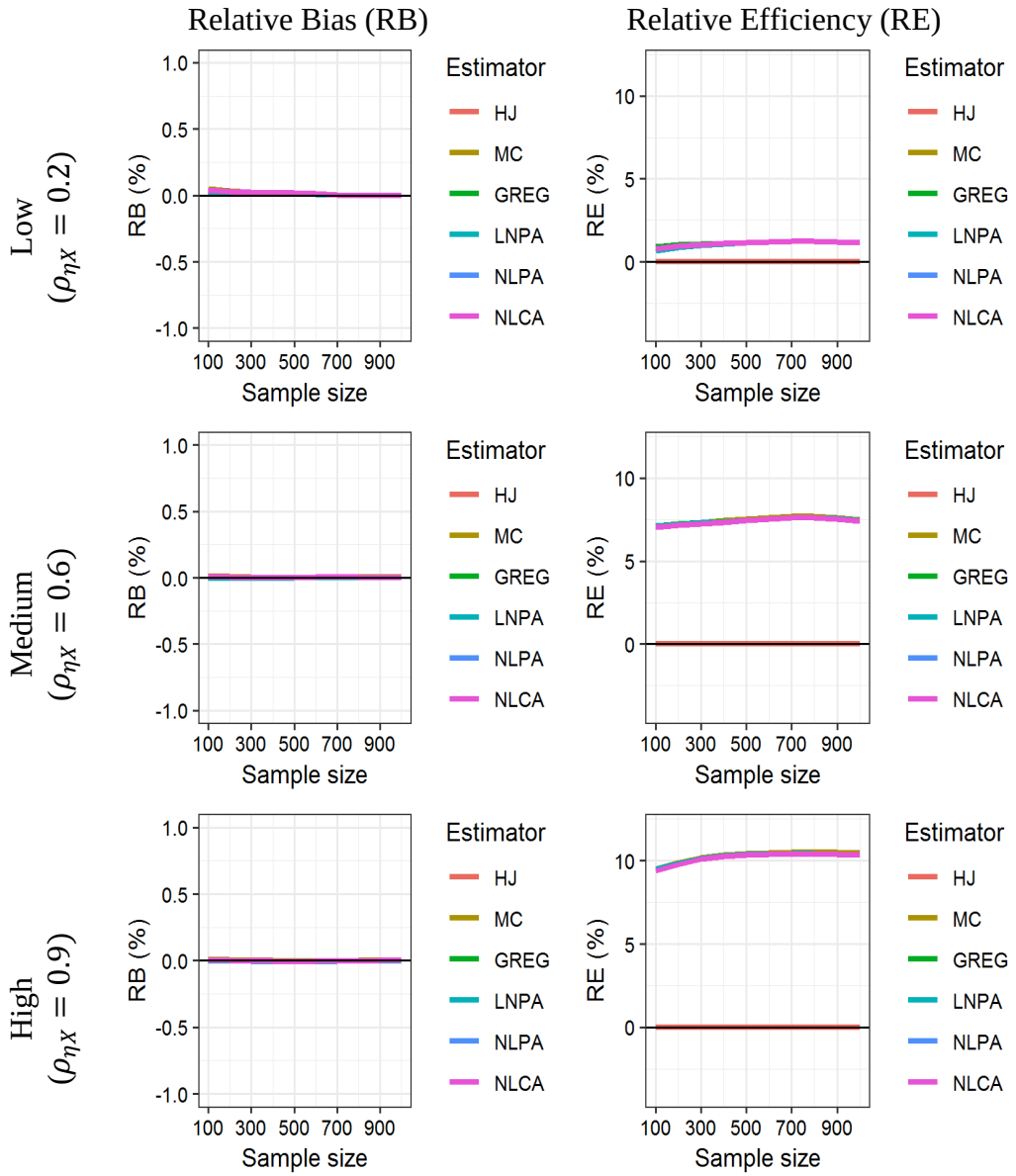


Figure A.2 Relative Bias (RB) and Relative efficiency (RE) of seven estimators as a function of the sample size for the population with a Bernoulli distribution with PPS designs.

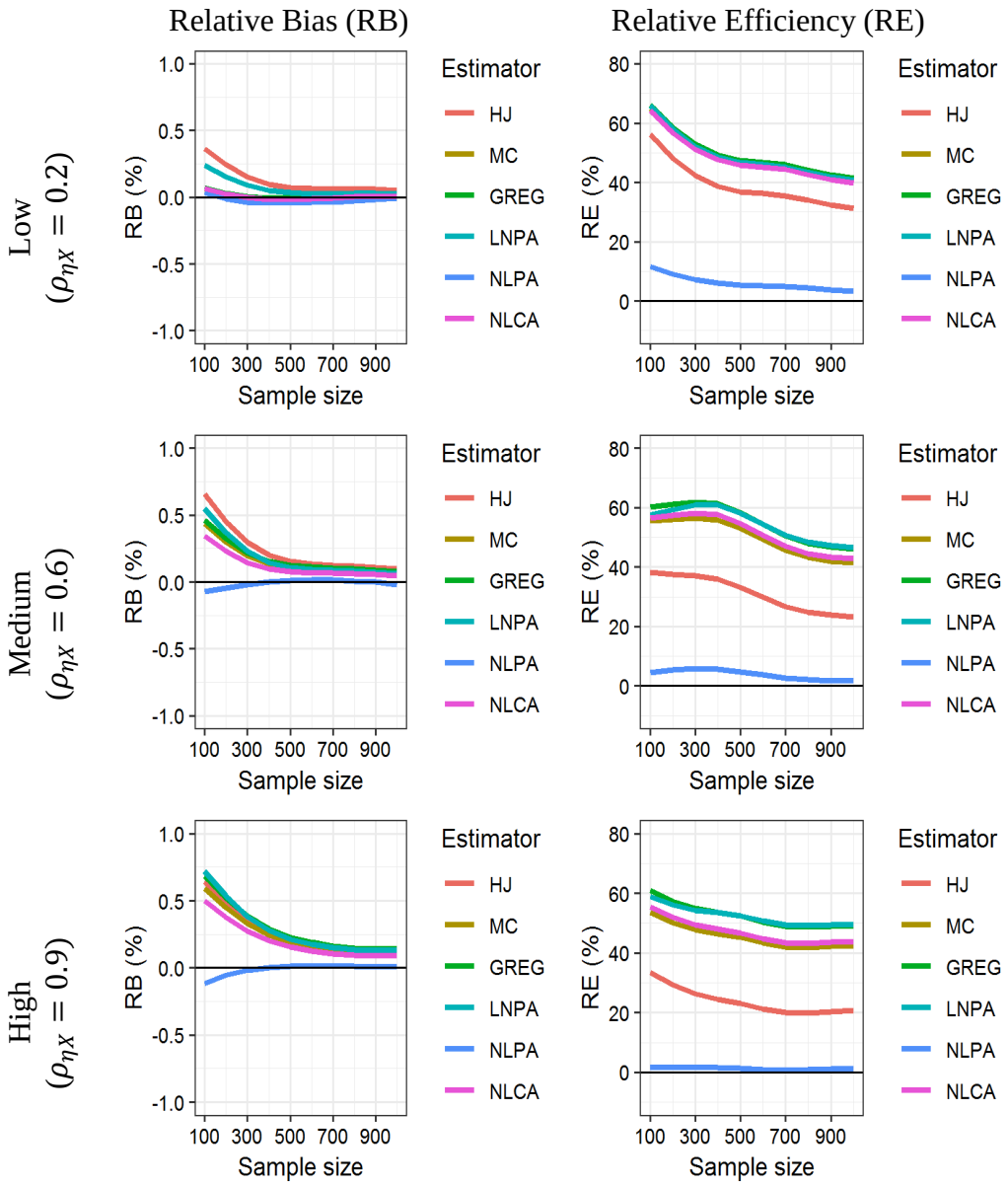


Figure A.3 Relative Bias (RB) and Relative efficiency (RE) of seven estimators as a function of the sample size for the population with a Bernoulli distribution with PO sampling designs.

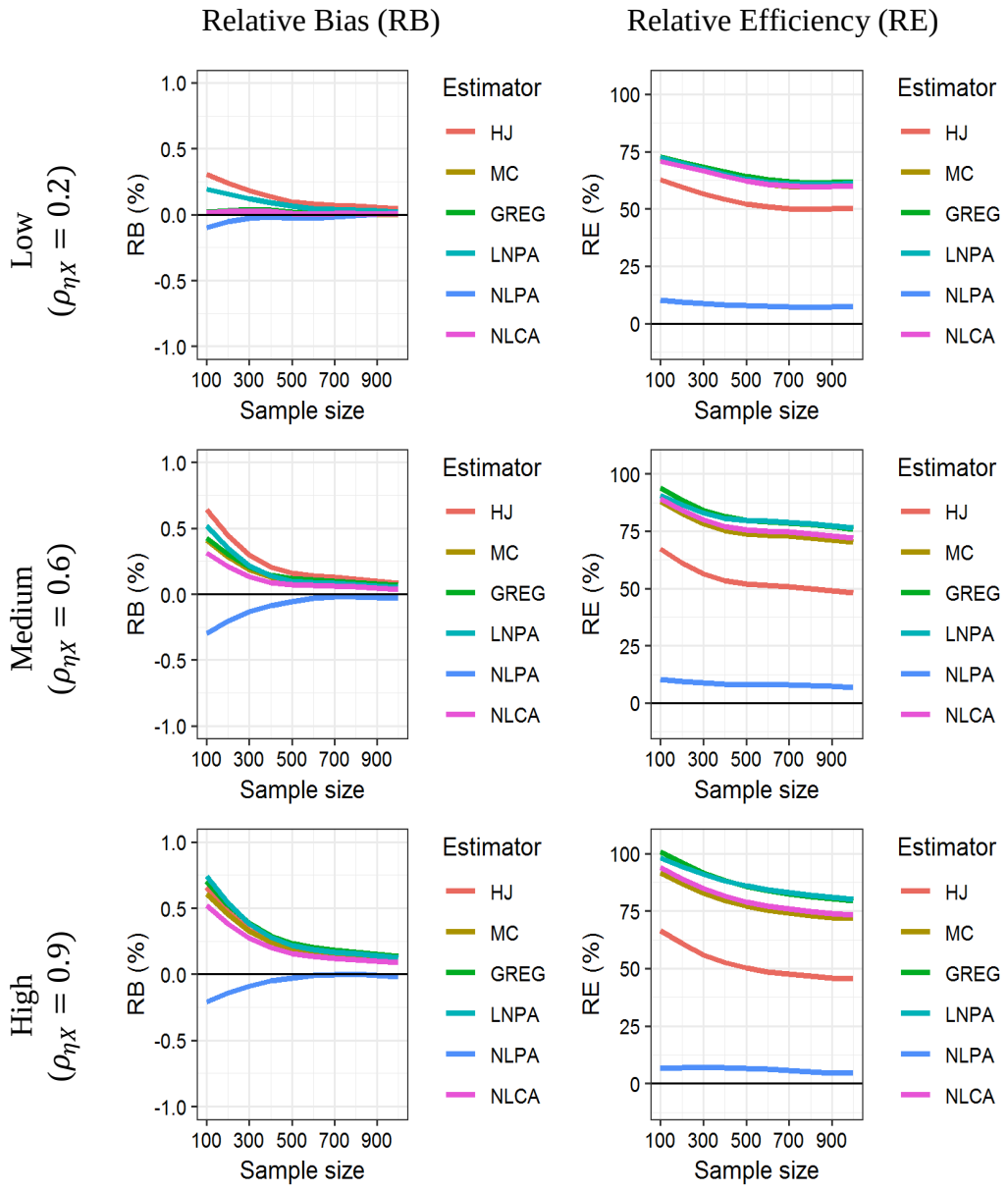


Figure A.4 Relative Bias (RB) and Relative efficiency (RE) of seven estimators as a function of the sample size for the population with a Poisson distribution with SRS designs.

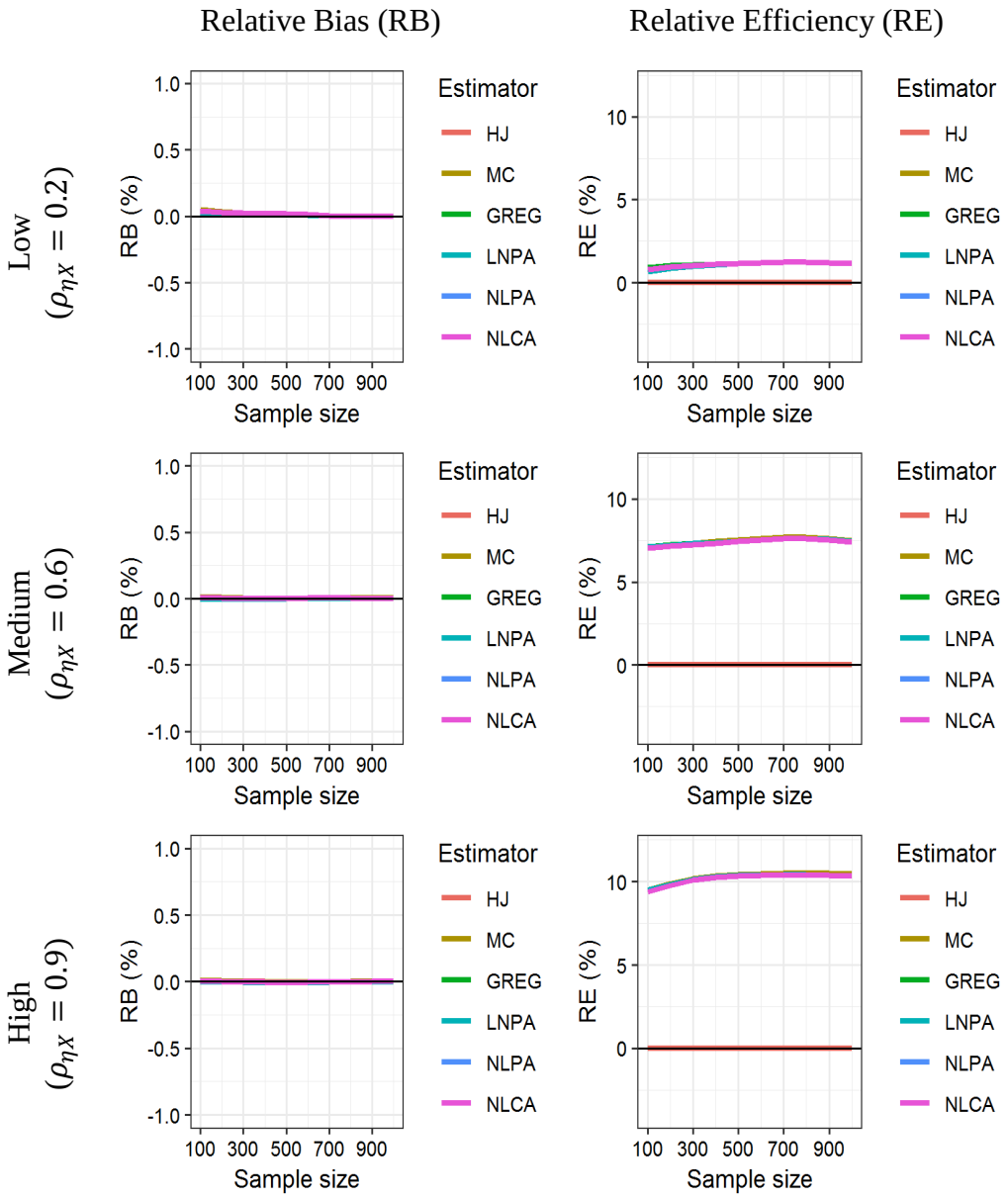


Figure A.5 Relative Bias (RB) and Relative efficiency (RE) of seven estimators as a function of the sample size for the population with a Poisson distribution with PPS designs.

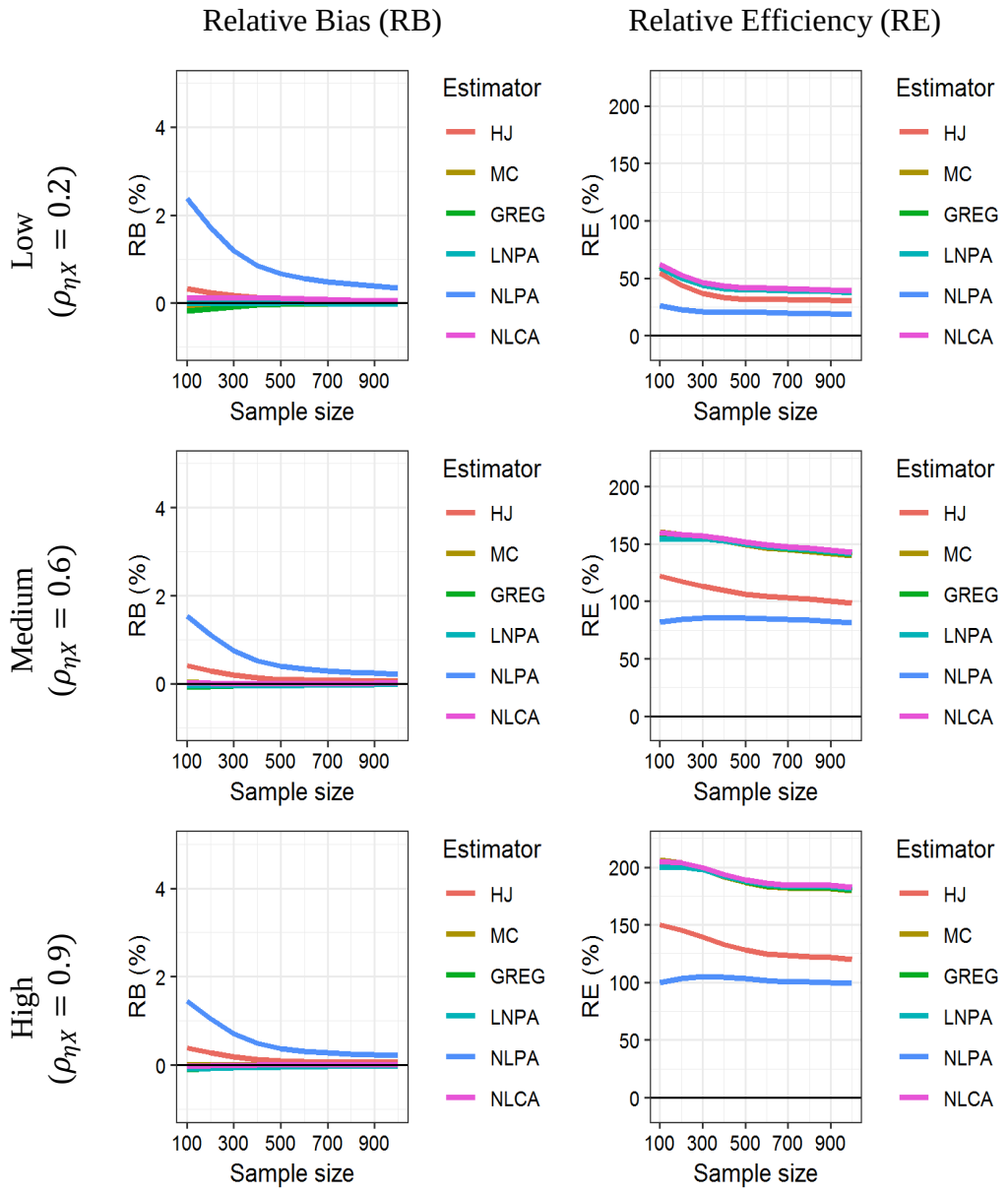


Figure A.6 Relative Bias (RB) and Relative efficiency (RE) of seven estimators as a function of the sample size for the population with a Poisson distribution with PO sampling designs.

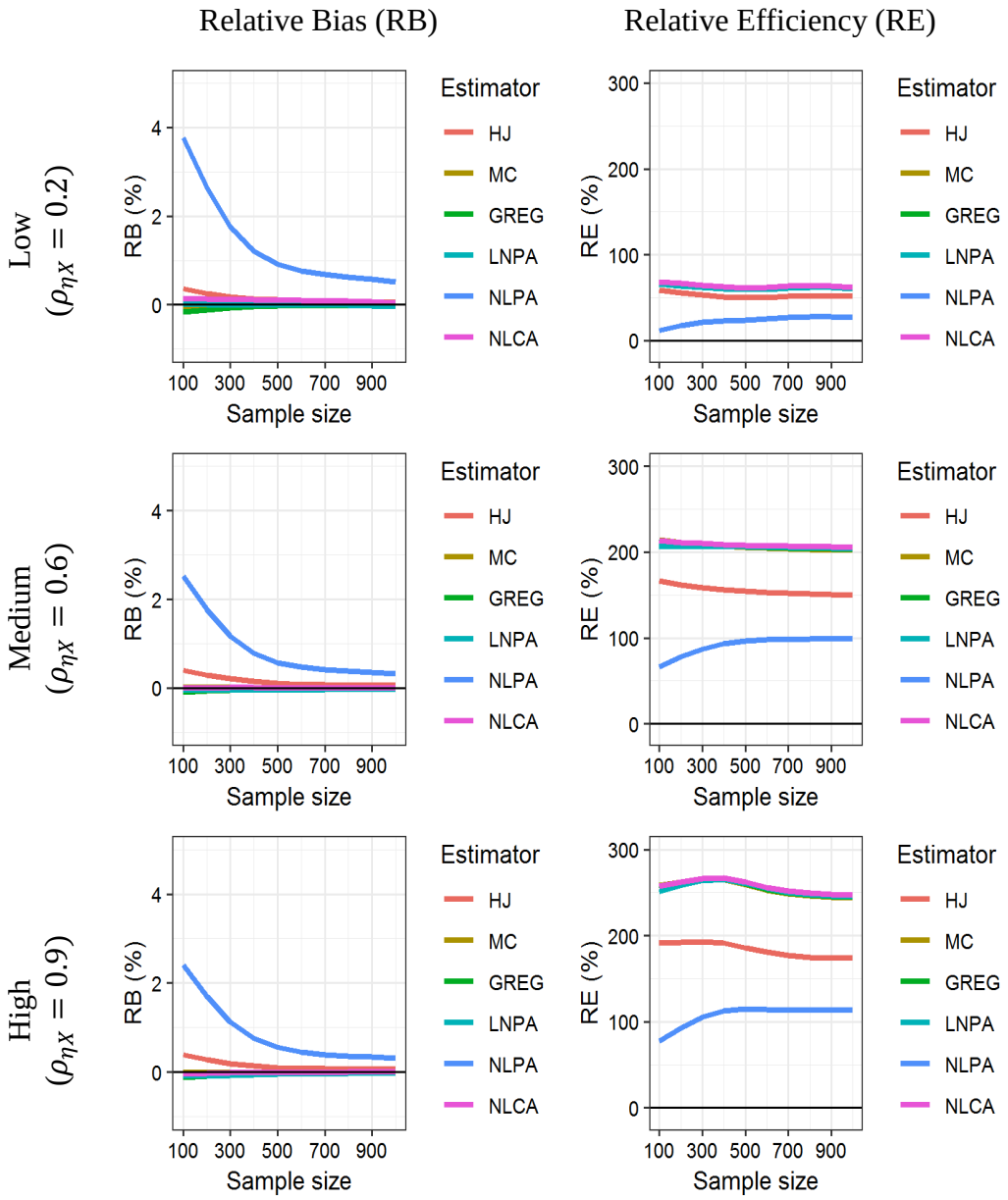


Figure A.7 Relative Bias (RB) and Relative efficiency (RE) of seven estimators as a function of the sample size for the population with a Gamma distribution with SRS designs.

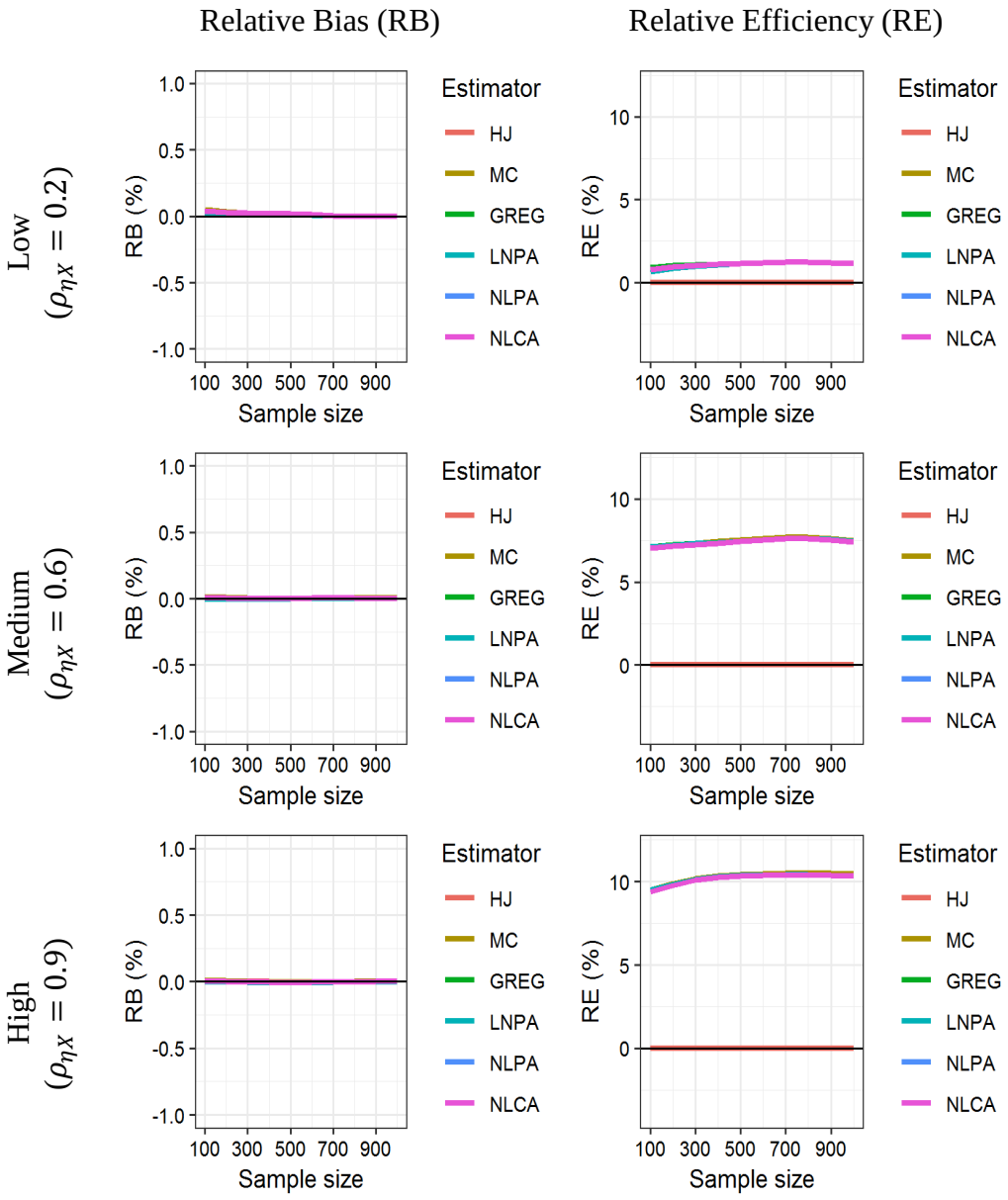


Figure A.8 Relative Bias (RB) and Relative efficiency (RE) of seven estimators as a function of the sample size for the population with a Gamma distribution with PPS designs.

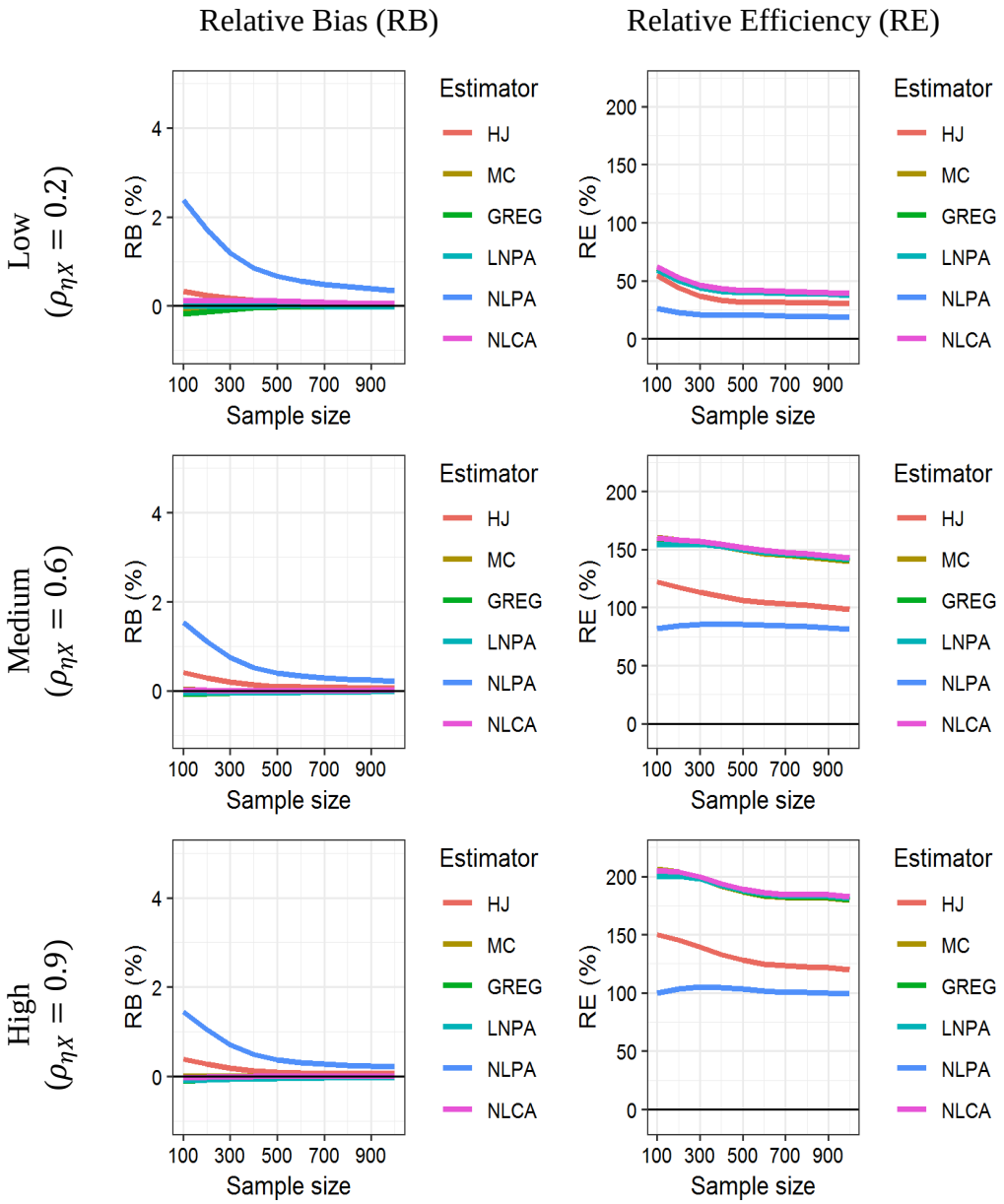
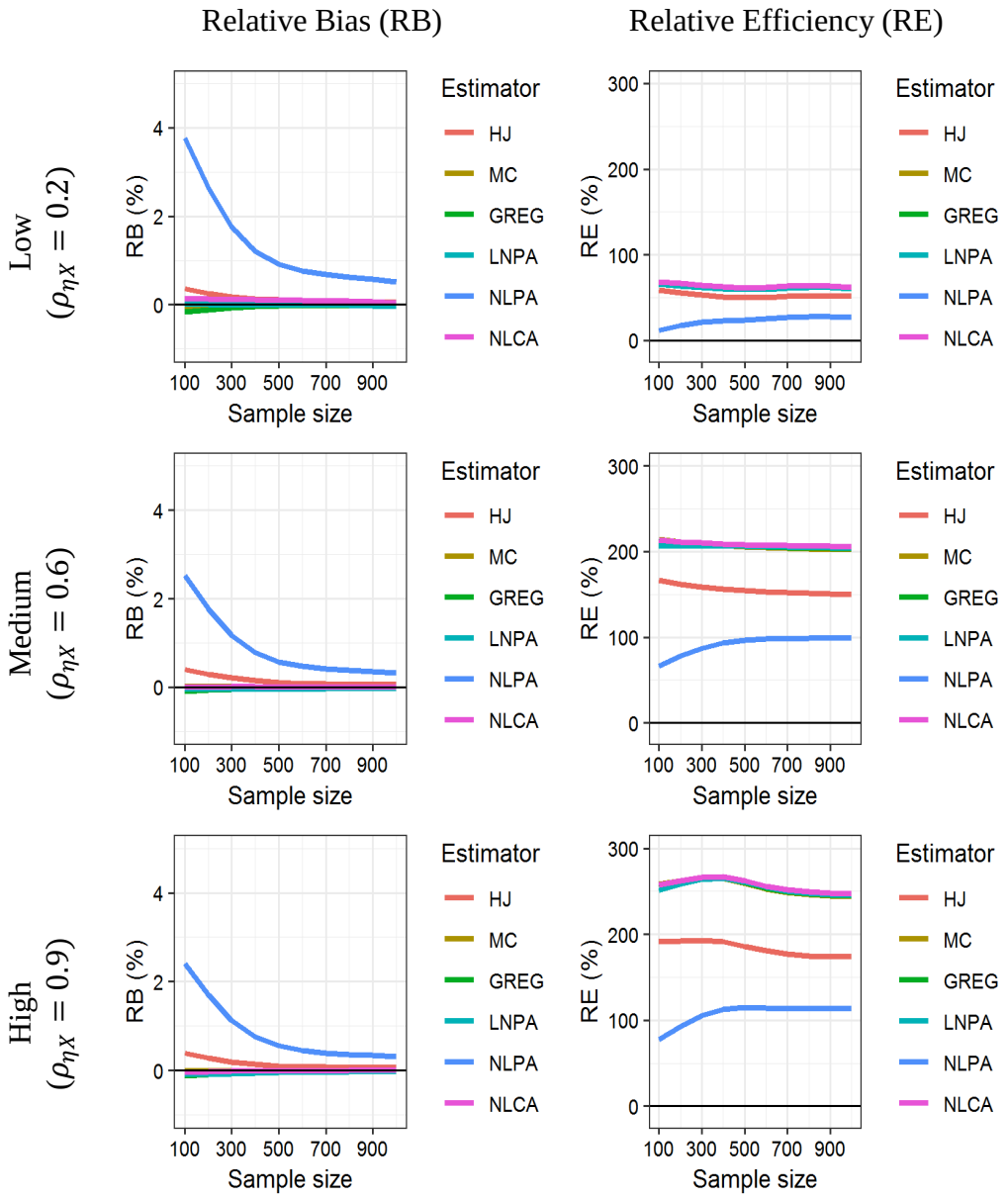


Figure A.9 Relative Bias (RB) and Relative efficiency (RE) of seven estimators as a function of the sample size for the population with a Gamma distribution with PO sampling designs.



A.2 Sample-Based AIC Estimator

Akaike (1981) defined "an information criterion" (AIC) as the estimator of $\mathbb{E}_y\mathbb{E}_x\left(\log\left(g\left(x\,|\,\hat{\theta}(y)\right)\right)\right)$ as

$$\text{AIC} = \widehat{AIC}_{mle} = -2\log\left(\mathcal{L}\left(\hat{\theta}\,|\,y\right)\right) + 2P\,, \tag{A.1}$$

where $\log\left(\mathcal{L}\left(\hat{\theta}\,|\,y\right)\right)$ is the numerical value of the log-likelihood at its maximum point, which corresponds to the values of the maximum likelihood estimates of θ, and P is the number of estimable parameters in the model. The latter term is a correction bias. The subscript mle indicates that the AIC is based on the MLE estimators.

We derive the sample-based AIC, dAIC as a plug-in estimator. Assume the function $\widehat{AIC}$ fitted to the population is sampled using a design defined by $\mathbf{S}$ such as $\mathbb{E}(\mathbf{S}) = \boldsymbol{\pi}$ and $\mathbb{V}(\mathbf{S}) = \boldsymbol{\Delta}$, then the sample-based estimator of $\widehat{AIC}_{mle}$ used in the PA framework is $dAIC$ defined as

$$dAIC = \widehat{AIC}_{pmle} = -2\sum_{k\in U} d_k S_k \log\left(\mathcal{L}_k\left(\hat{\theta}\,|\,y\right)\right) + 2P\,. \tag{A.2}$$

Equation (A.2) is the sample-based version of the AIC used in the PA approach. Although $\mathbb{E}\left(\sum_{k\in U} d_k S_k \log\left(\mathcal{L}_k\left(\hat{\theta}\,|\,y\right)\right)|\,\mathcal{F}\right) = \sum_{k\in U}\log\left(\mathcal{L}_k\left(\hat{\theta}\,|\,y\right)\right)$, there is no assurance that $\mathbb{E}\left(\widehat{AIC}_{pmle}\,|\,\mathcal{F}\right) = \widehat{AIC}_{mle}$ since $\mathbb{E}(P\,|\,\mathcal{F}) \neq P$ or the number of

parameters of the PML fitted to the sample, is an unbiased estimate of the number of parameters of the ML fitted to the entire population. Other alternatives address this problem but at the population level. One approach is the Takeuchi's Information Criterion (TIC, see Takeuchi 1976) which replaces P by $\mathrm{Tr}\left(\mathbf{J}\left(\hat{\theta}\right)\boldsymbol{\mathcal{I}}\left(\hat{\theta}\right)^{-1}\right)$.

The TIC is then an asymptotically unbiased estimate of the expected K-L information. However, Burnham & Anderson (2003) describe the problems with this approach since the estimation of the Jacobian $\mathbf{J}\left(\hat{\theta}\right)$ and Information matrix $\boldsymbol{\mathcal{I}}\left(\hat{\theta}\right)$ adjustment are computationally expensive and unstable in small samples. Lumley & Scott (2015) implements the AIC based on the TIC by replacing P by the sample-based estimate $\mathrm{Tr}\left(\mathbf{J}\left(\hat{\theta}\right)\boldsymbol{\mathcal{I}}\left(\hat{\theta}\right)^{-1}\right)$ in the instruction AIC from the R package survey (Lumley, 2012). Our experience confirmed the issues with this approach because this instruction computed imaginary values in the simulation runs.

We decided to use the number of parameters P in the PMLE because of the mathematical simplicity (i.e., count the number of parameters in the model). The reason being that it is unrealistic to assume that the PML model fitted to the sample can accommodate the same number of parameters as the ML model fitted to the population since the sample size is smaller, sometimes in several orders of magnitude than the population size. We do not expect to fit the same number of

parameters in the population model using a sample. The empirical results from the selection of variables based on the PA version of the AIC and the fact that the PA estimators perform slightly better than knowing the true model provide support for the use of this version of the AIC.

A.3 Theorems

A.3.1 Proof of Theorem 1.1

THEOREM 1.1 Assume a sequence of finite populations $\{\mathcal{F}_N\}_{N=1}^{\infty}$ of increasing size $U_N = \{1,...,N_N\}_{N=1}^{\infty}$ and samples $\{n_N\}_{N=1}^{\infty}$ drawn according to a sample design $\{p_N(A_N = a_N)\}_{N=1}^{\infty}$ satisfying the regularity conditions in Section 5.9 on page 252. The sequence of PA adjustment factors $\left\{\hat{\boldsymbol{\Gamma}}_{\mathbf{X},N}\right\}_{N=1}^{\infty}$ converges to the identity matrix $\mathbf{I} \in \mathbb{R}^{P\times P}$ as

$$\lim_{\substack{N\to\infty \\ n\to\infty}} \mathbb{E}\left(\hat{\boldsymbol{\Gamma}}_{\mathbf{X},N} - \mathbf{I} \mid \mathcal{F}\right) = \mathbf{0}.$$

We need to show that PA adjustment factor, $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}} = \mathbf{D}_{\mathbf{X}}\mathbf{D}_{\hat{\mathbf{X}}}^{-1} \in \mathbb{R}^{P\times P}$, is a design consistent estimator of the identity matrix $\mathbf{I} \in \mathbb{R}^{P\times P}$ where $\mathbf{D}_{\hat{\mathbf{X}}} = \text{diag}\left(\mathbf{d}^{\mathrm{T}}(\mathbf{x} \odot \mathbf{S})\right)$ is the diagonal matrix of the Horvitz-Thompson (HT) estimates of the auxiliary variables $\mathbf{x}_k = (x_{k1},\ldots,x_{kP})$ for $k \in U$,

$\mathbf{D}_{\mathbf{X}} = \text{diag}\left(\mathbf{1}^{\text{T}}\mathbf{X}\right)$ is the diagonal matrix of the auxiliary variable population totals $\mathbf{X} = \mathbf{1}^{\text{T}}\mathbf{x} = \left(X_1,...,X_P\right) \in \mathbb{R}^{1\times P}$, and $\mathbf{S} \in \{0,1\}$ is a discrete random variable for the design $p(\mathbf{S}=\mathbf{s})$ defined by $\mathbb{E}(\mathbf{S}) = \boldsymbol{\pi} \in (0,1)^{N\times 1}$ and $\mathbb{C}(\mathbf{S}) = \boldsymbol{\Delta} \in \mathbb{R}^{N\times N}$, and $\mathbf{d} = \boldsymbol{\pi}^{\odot -1} = [d_k] = [\pi_k] \in \mathbb{R}^{N\times 1}$ is the vector with the sampling weights defined as the inverse of the probabilities of inclusion.

Using the first two terms of the Taylor's Series expansion of the function $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}(\mathbf{S})$ evaluated at the point $\mathbf{S} = \boldsymbol{\pi}$, we can appropriate $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$ as

$$\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}(\mathbf{S}) = \hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\Big|_{\mathbf{S}=\boldsymbol{\pi}} + \frac{\partial \hat{\boldsymbol{\Gamma}}_{\mathbf{X}}}{\partial \mathbf{S}}\Big|_{\mathbf{S}=\boldsymbol{\pi}} (\mathbf{S}-\boldsymbol{\pi}) + \mathcal{O}_p\left(\|\mathbf{S}-\boldsymbol{\pi}\|_2^2\right). \tag{A.3}$$

To avoid tensor notation, we work on the elements of the diagonal $\hat{\Gamma}_{X_p}(\mathbf{S})$ using the alternative definition of $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$ as a diagonal matrix with the ratios of the population defined as

$$\left[\hat{\Gamma}_{X_{pq}}\right] = \begin{cases} \dfrac{X_{pq}}{\hat{X}_{pq}} & \text{If } p = q \\ 0 & \text{If } p \neq q \end{cases}. \tag{A.4}$$

for $p \neq q \in \{1,...,P\}$ .The first term is

$$\hat{\Gamma}_{X_p}\Big|_{\mathbf{S}=\boldsymbol{\pi}} = \frac{X_p}{\mathbf{d}^{\text{T}}\left(X_p \odot \mathbf{S}\right)}\Big|_{\mathbf{S}=\boldsymbol{\pi}} = \left[\frac{X_p}{\mathbf{d}^{\text{T}}\left(\mathbf{x}_p \odot \boldsymbol{\pi}\right)}\right] = \left[\frac{X_p}{X_p}\right] = 1.$$

Working on the second term with the partial evaluated at $\mathbf{S} = \boldsymbol{\pi}$:

$$\left.\frac{\partial \hat{\boldsymbol{\Gamma}}_{X_p}}{\partial \mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}} = \left.\frac{\partial}{\partial \mathbf{S}}\left(\frac{X_p}{\mathbf{d}^{\mathrm{T}}\left(\mathbf{x}_p \odot \mathbf{S}\right)}\right)\right|_{\mathbf{S}=\boldsymbol{\pi}} = \left.-\frac{X_p\left(\mathbf{d}\odot\mathbf{x}_p\right)^{\mathrm{T}}}{\left(\mathbf{d}^{\mathrm{T}}\left(\mathbf{x}_p \odot \mathbf{S}\right)\right)^2}\right|_{\mathbf{S}=\boldsymbol{\pi}}$$

$$= -\frac{X_p\left(\mathbf{d}\odot\mathbf{x}_p\right)^{\mathrm{T}}}{\left(\mathbf{d}^{\mathrm{T}}\left(\mathbf{x}_p \odot \boldsymbol{\pi}\right)\right)^2} = -\frac{X_p}{X_p^2}\left(\mathbf{d}\odot\mathbf{x}_k\right)^{\mathrm{T}} \quad . \qquad \text{(A.5)}$$

$$= -\frac{\left(\mathbf{d}\odot\mathbf{x}_p\right)^{\mathrm{T}}}{X_p} \in \mathbb{R}^{1\times N}$$

Condition (a): The estimator $\hat{\Gamma}_{X_p}$ is asymptotically unbiased for 1 as

$$\mathbb{E}\left(\hat{\Gamma}_{X_p}\right) = \mathbb{E}\left(1-\left(\mathbf{d}\odot x_p\right)^{\mathbf{T}}\left(\mathbf{S}-\boldsymbol{\pi}\right)+O\left(\frac{1}{N}\right)\right), \qquad \text{(A.6)}$$
$$= 1+O\left(\frac{1}{N}\right)$$

where it is the same for all elements of the diagonal, so $\mathbb{E}\left(\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\right) = \mathbf{I}+O\left(\frac{1}{N}\right)$.

Condition (b): The variance of estimator $\frac{1}{N}\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$ goes to zero as $N\to\infty$. We begin by rewriting the variance of $\hat{\Gamma}_{X_p}$ as a function of the variance of $\hat{\bar{X}}_{HT,p}$ as

$$\mathbb{V}\left(\frac{\hat{\Gamma}_{X_p}}{N}\right) = \mathbb{V}\left(\frac{\left(\mathbf{d}\odot x_p\right)^{\mathbf{T}}}{NX_p}\left(\mathbf{S}-\boldsymbol{\pi}\right)\right)+O\left(\frac{1}{N}\right). \qquad \text{(A.7)}$$
$$= \frac{1}{X_p^2}\mathbb{V}\left(\hat{\bar{X}}_{HT,p}\right)+O\left(\frac{1}{N}\right)$$

Since we already proved that $\mathbb{V}\left(\hat{\bar{X}}_{HT,p}\right)$ goes to zero as $N \to \infty$, then the same applies to the totals in diagonals. Since conditions (a) and (b) are met, then the sequence of estimators $\left\{\hat{\boldsymbol{\Gamma}}_{N,\mathbf{X}}\right\}_{N=1}^{\infty}$ is a design consistent estimator of $\mathbf{I}_N$.

The approximate variance-covariance of $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$, $\mathbb{C}\left(\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\right)$, is

$$\mathbb{AC}\left(\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\right) = diag\left(\frac{\mathbb{AV}\left(\hat{X}_{HT,1}\right)}{X_1^2},...,\frac{\mathbb{AV}\left(\hat{X}_{HT,P}\right)}{X_P^2}\right), \tag{A.8}$$

where $\mathbb{AV}\left(\hat{X}_{HT,p}\right)$ is the approximate variance of the HT total of the auxiliary variable x_p computed as $\mathbb{AV}\left(\hat{X}_{HT,p}\right) = \left(\mathbf{d} \odot x_p\right)^{\mathrm{T}} \boldsymbol{\Delta}\left(\mathbf{d} \odot x_p\right)$ for $p \in \{1,...,P\}$.

The variance-covariance estimator of $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$, $\hat{\mathbb{C}}\left(\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\right) \in \mathbb{R}^{P\times P}$, is

$$\hat{\mathbb{C}}\left(\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\right) = diag\left(\frac{\hat{\mathbb{V}}\left(\hat{X}_{HT,1}\right)}{X_1^2},...,\frac{\hat{\mathbb{V}}\left(\hat{X}_{HT,P}\right)}{X_P^2}\right), \tag{A.9}$$

where $\hat{\mathbb{V}}\left(\hat{X}_{HT,p}\right) = \left(\mathbf{d} \odot x_p\right)^{\mathrm{T}} \hat{\boldsymbol{\Delta}}\left(\mathbf{d} \odot x_p\right)$ and $\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$.

A.3.2 Variance-Covariance of $\hat{\boldsymbol{\beta}}_{pmle}$ in a Normal Linear Model

Let y the variable of interest with a superpopulation model $\mathcal{M}_y$ where $y_k \sim \mathcal{N}\left(\mathbf{x}_k\boldsymbol{\beta},\sigma^2\right)$, $\mathbf{x}_k = \left(x_{k1},\ldots,x_{kP}\right) \in \mathbb{R}^{1\times P}$ is the vector of auxiliary variables and $\boldsymbol{\beta} = \left(\beta_{k1},\ldots,\beta_{kP}\right)^{\mathrm{T}} \in \mathbb{R}^{P\times 1}$ is the vector with the location parameters. Let $\mathcal{F}$

be a finite population consisting of N *id* realizations of $\mathcal{M}_y$. Let $\mathbf{S}$ be a random discrete vector that defines the sample design $p(\mathbf{S}=\mathbf{s})$ with $\mathbb{E}(\mathbf{S})=\boldsymbol{\pi}$ and $\mathbb{C}(\mathbf{S})=\boldsymbol{\Delta}$ that meets the regularity conditions listed in Section 5.9. Assume that a normal PL model is fitted to the sample. The vector of the PMLE estimators $\hat{\boldsymbol{\beta}}_{mle} \in \mathbb{R}^{P\times 1}$ is

$$\hat{\boldsymbol{\beta}}_{pmle} = \hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}}, \tag{A.10}$$

where

$$\hat{\mathbf{T}}_{\mathbf{xx}} = \left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{x} = \left[\sum_{k\in U} d_k S_k x_{ik} x_{jk}\right] \in \mathbb{R}^{P\times P}, \text{ and} \tag{A.11}$$

$$\hat{\mathbf{T}}_{\mathbf{xy}} = \left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{y} = \left[\sum\nolimits_{k\in U} x_{ik} y_{jk}\right] \in \mathbb{R}^{P\times 1}. \tag{A.12}$$

See Binder (1983) for the proof that $\hat{\boldsymbol{\beta}}_{pmle}$ is a design consistent estimator of $\hat{\boldsymbol{\beta}}_{mle}$, that is

$$\begin{gathered}\lim_{N\to\infty} \mathbb{E}\left(\hat{\boldsymbol{\beta}}_{mle} - \hat{\boldsymbol{\beta}}_{pmle}\right) = \mathbf{0} \in \mathbb{R}^{P\times 1}, \text{ and} \\ \lim_{N\to\infty} \mathbb{C}\left(\hat{\boldsymbol{\beta}}_{pmle}\right) = \mathbf{0} \in \mathbb{R}^{P\times P}\end{gathered} \tag{A.13}$$

.

The variance-covariance $\mathbb{C}\left(\hat{\boldsymbol{\beta}}_{pmle}\right)$ is computed using the first two terms of the TS approximation of the function $\hat{\boldsymbol{\beta}}_{pmle}(\mathbf{S})$ evaluated at the point $\mathbf{S}=\boldsymbol{\pi}$ as

$$\hat{\boldsymbol{\beta}}_{pmle}(\mathbf{S}) = \hat{\boldsymbol{\beta}}_{pmle}(\mathbf{S})\Big|_{\mathbf{S}=\boldsymbol{\pi}} + \left(\frac{\partial \hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}}}{\partial \mathbf{S}}\right)^{\mathrm{T}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}} (\mathbf{S}-\boldsymbol{\pi}) + \mathcal{O}_p\left(\|\mathbf{S}-\boldsymbol{\pi}\|_2^2\right). \tag{A.14}$$

Working on the first term,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{pmle}\Big|_{\mathbf{S}=\boldsymbol{\pi}} &= \hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}}\Big|_{\mathbf{S}=\boldsymbol{\pi}} \\ &= \left(\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{x}\right)^{-1}\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{y}\Big|_{\mathbf{S}=\boldsymbol{\pi}}. \\ &= \mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{T}_{\mathbf{xy}} \\ &= \hat{\boldsymbol{\beta}}_{mle}\end{aligned} \tag{A.15}$$

Working on the second term of (A.2) and using the chain rule for derivatives of matrices

$$\begin{aligned}\left.\frac{\partial\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}}}{\partial\mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}} &= \frac{\partial\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}}{\partial\mathbf{S}}\hat{\mathbf{T}}_{\mathbf{xy}} + \hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\frac{\partial\hat{\mathbf{T}}_{\mathbf{xy}}}{\partial\mathbf{S}}. \\ &= \mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2\end{aligned} \tag{A.16}$$

The partial derivative of a matrix with respect to the vector **S** generates a 3-dimensional matrix of size $P\times P\times N$. We will not introduce vector notation since the matrix becomes of size $P\times P$.

$$\begin{aligned}\mathbf{A}_1 &= \hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\left(\frac{\partial\hat{\mathbf{T}}_{\mathbf{xx}}}{\partial\mathbf{S}}\right)\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}}\Big|_{\mathbf{S}=\boldsymbol{\pi}} \\ &= -\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\frac{\partial\left(\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{x}\right)}{\partial\mathbf{S}}\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}}\Big|_{\mathbf{S}=\boldsymbol{\pi}}. \\ &= -\mathbf{T}_{\mathbf{x},\mathbf{x}}^{-1}\begin{bmatrix}\mathbf{x}_1\odot\mathbf{d} \\ \cdots \\ \mathbf{x}_P\odot\mathbf{d}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_P\end{bmatrix}\hat{\boldsymbol{\beta}}_{mle}\end{aligned} \tag{A.17}$$

Computing the second term $\mathbf{A}_2 = \hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\left.\frac{\partial\hat{\mathbf{T}}_{\mathbf{xy}}}{\partial\mathbf{S}}\right|_{\mathbf{S}=\boldsymbol{\pi}}$ from (A.17),

$$\mathbf{A}_2 = \hat{\mathbf{T}}_{\mathbf{xx}}^{-1} \left. \frac{\partial \left(\mathbf{S} \odot \mathbf{d} \odot \mathbf{x}\right)^{\mathrm{T}} \mathbf{y}}{\partial \mathbf{S}} \right|_{\mathbf{S}=\boldsymbol{\pi}} = \mathbf{T}_{\mathbf{xx}}^{-1} \begin{bmatrix} \mathbf{x}_1 \odot \mathbf{d} \\ \cdots \\ \mathbf{x}_P \odot \mathbf{d} \end{bmatrix} \mathbf{y} \quad . \tag{A.18}$$

Putting terms $\mathbf{A}_1$ and $\mathbf{A}_2$ we obtain

$$\mathbf{A} = -\mathbf{T}_{\mathbf{xx}}^{-1} \begin{bmatrix} \mathbf{x}_1 \odot \mathbf{d} \\ \cdots \\ \mathbf{x}_P \odot \mathbf{d} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_P \end{bmatrix} \hat{\boldsymbol{\beta}}_{mle} + \mathbf{T}_{\mathbf{xx}}^{-1} \begin{bmatrix} \mathbf{x}_1 \odot \mathbf{d} \\ \cdots \\ \mathbf{x}_P \odot \mathbf{d} \end{bmatrix} \mathbf{y} = \mathbf{T}_{\mathbf{xx}}^{-1} \begin{bmatrix} \mathbf{x}_1 \odot \mathbf{d} \odot \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{mle}\right) \\ \cdots \\ \mathbf{x}_P \odot \mathbf{d} \odot \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{mle}\right) \end{bmatrix} , \tag{A.19}$$

where $\mathbf{e} = \mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{mle}$ is the vector with the residuals of the ML model fitted to the population. The approximate variance-covariance is obtained computing the variance of $\mathbf{A}$ as

$$\mathbb{AC}\left(\hat{\boldsymbol{\beta}}_{pmle}\right) = \mathbf{T}_{\mathbf{xx}}^{-1} \mathbb{AC} \left(\begin{bmatrix} \mathbf{x}_1 \odot \mathbf{d} \odot \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{mle}\right) \\ \cdots \\ \mathbf{x}_P \odot \mathbf{d} \odot \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{mle}\right) \end{bmatrix} \right) \mathbf{T}_{\mathbf{xx}}^{-1} = \mathbf{T}_{\mathbf{xx}}^{-1} \mathbb{AC} \left(\begin{bmatrix} \mathbf{x}_1 \odot \mathbf{d} \odot \mathbf{e} \\ \cdots \\ \mathbf{x}_P \odot \mathbf{d} \odot \mathbf{e} \end{bmatrix} \right) \mathbf{T}_{\mathbf{xx}}^{-1} \quad . \tag{A.20}$$

$$= \mathbf{T}_{\mathbf{xx}}^{-1} \begin{pmatrix} V_{11} & \cdots & V_{1P} \\ \cdots & \cdots & \cdots \\ V_{P1} & \cdots & V_{1P} \end{pmatrix} \mathbf{T}_{\mathbf{xx}}^{-1}$$

where $V_{pq} = \left(\mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e}\right)^{\mathrm{T}} \boldsymbol{\Delta} \left(\mathbf{x}_q \odot \mathbf{d} \odot \mathbf{e}\right)$. The approximate variance-covariance between $\hat{\beta}_{pmle,p}$ and $\hat{\beta}_{pmle,p}$ is

$$\mathbb{AC}\left(\hat{\beta}_{pmle,p}, \hat{\beta}_{pmle,q}\right) = T_{x_p x_q}^{-1} \left(x_p \odot \mathbf{d} \odot \mathbf{e}\right)^{\mathrm{T}} \boldsymbol{\Delta} \left(x_q \odot \mathbf{d} \odot \mathbf{e}\right) T_{x_p x_q}^{-1} . \quad \text{(A.21)}$$

The variance estimator $\hat{\mathbb{C}}\left(\hat{\beta}_{pmle,p}, \hat{\beta}_{pmle,q}\right)$, computed by replacing the unknown population quantities by their sample-based estimates, is

$$\hat{\mathbb{C}}\left(\hat{\beta}_{pmle,p}, \hat{\beta}_{pmle,q}\right) = \hat{T}_{x_p x_q}^{-1} \left(x_p \odot \mathbf{d} \odot \breve{\mathbf{e}}\right)^{\mathrm{T}} \hat{\boldsymbol{\Delta}} \left(x_q \odot \mathbf{d} \odot \breve{\mathbf{e}}\right) \hat{T}_{x_p x_q}^{-1} , \quad \text{(A.22)}$$

where $\breve{\mathbf{e}} = \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right) \odot \mathbf{s} \oslash \boldsymbol{\pi}$ is the vector with the sample-based residuals of the PL model, $\hat{T}_{x_p x_q}$ is the element (p,q) of $\hat{\mathbf{T}}_{\mathbf{xx}}$, the matrix of the HT estimates of the population of the cross product $\mathbf{x}^{\mathrm{T}}\mathbf{x}$, and $\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$. The expression (A.22) matches those found in Binder (1983), Särndal, Swensson, & Wretman (1992), and Fuller (2009).

A.3.3 Variance-Covariance of $\hat{\boldsymbol{\beta}}_{pa}$ in a Normal Linear Model

Let y the variable of interest with a superpopulation model $\mathcal{M}_y$ where $y_k \sim \mathcal{N}\left(\mathbf{x}_k\boldsymbol{\beta}, \sigma^2\right)$, $\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kP}\right) \in \mathbb{R}^{1\times P}$ is the vector of auxiliary variables and $\boldsymbol{\beta} = \left(\beta_{k1}, \ldots, \beta_{kP}\right)^{\mathrm{T}} \in \mathbb{R}^{P\times 1}$ is the vector with the location parameters. Let $\mathcal{F}$ be a finite population consisting of N *iid* realizations of $\mathcal{M}_y$. Let $\mathbf{S}$ be a random discrete vector that uniquely defines the sample design $p\left(\mathbf{S} = \mathbf{s}\right)$ with $\mathbb{E}\left(\mathbf{S}\right) = \boldsymbol{\pi}$

and $\mathbb{C}(\mathbf{S}) = \boldsymbol{\Delta}$ that meets the regularity conditions listed in Section 5.9. Assume that a normal PL model is fitted to the sample. The PA estimator of $\hat{\boldsymbol{\beta}}_{mle} \in \mathbb{R}^{P\times 1}$ is

$$\hat{\boldsymbol{\beta}}_{pa} = \hat{\boldsymbol{\Gamma}}_{\mathbf{X}} \hat{\boldsymbol{\beta}}_{pmle}, \tag{A.23}$$

where $\hat{\boldsymbol{\beta}}_{pmle}$ is the vector with the PML estimates of $\hat{\boldsymbol{\beta}}_{mle} \in \mathbb{R}^{P\times 1}$ described in Section A.3.2 and $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$ is the PA adjustment matrix described in Section A.3.3.

The sequence of PA estimators $\left\{\hat{\boldsymbol{\beta}}_{pa,N}\right\}_{N=1}^{\infty}$ is design consistent of $\hat{\boldsymbol{\beta}}_{mle,N}$ since it is the product of the sequence of estimates $\left\{\hat{\boldsymbol{\beta}}_{pmle,N}\right\}_{N=1}^{\infty}$, which is design consistent of $\hat{\boldsymbol{\beta}}_{mle,N}$ (see Binder, 1983), and the sequence of PA adjustments $\left\{\hat{\boldsymbol{\Gamma}}_{\mathbf{X},N}\right\}_{N=1}^{\infty}$, which is design consistent of the identity matrix $\mathbf{I}_N$ after applying Slutsky's theorem. In other words, the following two conditions hold

$$\begin{gathered}\lim_{N\to\infty} \mathbb{E}\left(\hat{\boldsymbol{\beta}}_{mle} - \hat{\boldsymbol{\beta}}_{pa}\right) = \mathbf{0} \in \mathbb{R}^{P\times 1} \text{ and} \\ \lim_{N\to\infty} \mathbb{C}\left(\hat{\boldsymbol{\beta}}_{pa}\right) = \mathbf{0} \in \mathbb{R}^{P\times P}.\end{gathered} \tag{A.24}$$

The approximate variance-covariance $\mathbb{C}\left(\hat{\boldsymbol{\beta}}_{pa}\right)$ is computed using the first two terms of the TS approximation of the function $\hat{\boldsymbol{\beta}}_{pa}(\mathbf{S})$ evaluated at the point $\mathbf{S} = \boldsymbol{\pi}$ as

$$\hat{\boldsymbol{\beta}}_{pa}(\mathbf{S}) = \hat{\boldsymbol{\beta}}_{pa}(\mathbf{S})\Big|_{\mathbf{S}=\boldsymbol{\pi}} + \left(\frac{\partial \hat{\boldsymbol{\beta}}_{pa}}{\partial \mathbf{S}}\right)^{\mathrm{T}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}} (\mathbf{S}-\boldsymbol{\pi}) + \mathcal{O}_p\left(\|\mathbf{S}-\boldsymbol{\pi}\|_2^2\right). \quad \text{(A.25)}$$

Working on the term $\left(\frac{\partial \hat{\boldsymbol{\beta}}_{pa}}{\partial \mathbf{S}}\right)^{\mathrm{T}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}}$

$$\begin{aligned}\frac{\partial \hat{\boldsymbol{\beta}}_{pa}}{\partial \mathbf{S}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}} &= \frac{\partial \hat{\boldsymbol{\Gamma}}_{\mathbf{X}} \hat{\boldsymbol{\beta}}_{pmle}}{\partial \mathbf{S}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}} \\ &= \frac{\partial \hat{\boldsymbol{\Gamma}}_{\mathbf{X}}}{\partial \mathbf{S}} \hat{\boldsymbol{\beta}}_{pmle}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}} + \hat{\boldsymbol{\Gamma}}_{\mathbf{X}} \frac{\partial \hat{\boldsymbol{\beta}}_{pmle}}{\partial \mathbf{S}}\Bigg|_{\mathbf{S}=\boldsymbol{\pi}}\end{aligned}. \quad \text{(A.26)}$$

The partial derivatives $\frac{\partial \hat{\boldsymbol{\Gamma}}_{\mathbf{X}}}{\partial \mathbf{S}}$ $\frac{\partial \hat{\boldsymbol{\beta}}_{pmle}}{\partial \mathbf{S}}$ were derived in Sections A.3.1 and A.3.2. Combing these results, the approximate variance-covariance between $\hat{\beta}_{pa,p}$ and $\hat{\beta}_{pa,q}$ is

$$\begin{aligned}\mathbb{AC}\left(\hat{\beta}_{pa,p}, \hat{\beta}_{pa,q}\right) = \mathbf{T}^{-1}_{\mathbf{x}_p\mathbf{x}_q}\left[V_{pq}\right]\mathbf{T}^{-1}_{\mathbf{x}_p\mathbf{x}_q} + \left[W_{pq}\right] \\ +2\mathbf{T}^{-1}_{\mathbf{x}_p\mathbf{x}_q}\left[VW_{pq}\right]\mathbf{T}^{-1}_{\mathbf{x}_p\mathbf{x}_q}\end{aligned}, \quad \text{(A.27)}$$

where

- $V_{pq} = \left(\mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e}\right)^{\mathrm{T}} \boldsymbol{\Delta} \left(\mathbf{x}_q \odot \mathbf{d} \odot \mathbf{e}\right)$ is the contribution to the variance form fitting the PL model $\mathbf{y} = \hat{\boldsymbol{\beta}}_{pmle}\mathbf{x}$ with residuals $\mathbf{e} = \mathbf{y} - \hat{\boldsymbol{\beta}}_{pmle}\mathbf{x}$.

- $W_{pp} = \left(\frac{\hat{\beta}_{mle,p}}{X_p} \mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e} \right)^{\mathrm{T}} \boldsymbol{\Delta} \left(\frac{\hat{\beta}_{mle,p}}{X_p} \mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e} \right)$ or $W_{pq} = 0$ if $p \neq q \in \{1,...,P\}$ is the contribution to the variance form the PA adjustment $\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}$.

- $VW_p = 2 \left(\frac{\hat{\beta}_{mle,p}}{X_p} \mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e} \right)^{\mathrm{T}} \boldsymbol{\Delta} \left(\mathbf{x}_p \odot \mathbf{d} \odot \mathbf{e} \right)$ or $VW_{pq} = 0$ if $p \neq q \in \{1,...,P\}$ is the contribution to the variance form the covariance between the PA adjustment and the PL model $\mathbf{y} = \hat{\boldsymbol{\beta}}_{pmle} \mathbf{x}$.

The variance estimator $\hat{\mathbb{C}}\left(\hat{\beta}_{pa,p}, \hat{\beta}_{pa,q}\right)$, computed by replacing the unknown population quantities by their sample-based estimates, that is, $\mathbf{e} = \mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{mle}$ by $\breve{\mathbf{e}} = \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right) \odot \mathbf{s}$, the elements (p,q) of $\mathbf{T_{xx}}$ by $\hat{\mathbf{T}}_{\mathbf{xx}}$, the matrix of the HT estimates of the population of the cross product totals of $\mathbf{x}$, and $\boldsymbol{\Delta}$ by $\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta} \oslash \boldsymbol{\Pi}$.

A.4 Empirical Summary Measures Used in Monte Carlo Simulations

The summary measures for bias and accuracy for Monte Carlo Simulations for a fixed population $\mathcal{F}$ are defined as

$$RB\left(\hat{Y}_E\right)\% = 100 \times \frac{1}{B} \sum_{b=1}^{B} \frac{\hat{Y}_{E,b} - Y}{Y}, \qquad \text{(A.28)}$$

$$MSE\left(\hat{Y}_E\right) = \frac{\sum_{b=1}^{B}\left(\hat{Y}_{E,b} - Y\right)^2}{B}, \tag{A.29}$$

$$RRMSE = \sqrt{\frac{MSE\left(\hat{Y}_E\right)}{Y^2}}, \text{ and} \tag{A.30}$$

$$RE\left(\hat{Y}_E\right)\% = 100 \times \left(\frac{MSE\left(\hat{Y}_{HT}\right)}{MSE\left(\hat{Y}_E\right)} - 1 \right), \tag{A.31}$$

where $\hat{Y}_E$ is the estimator being evaluated and $\hat{Y}_{E,b}$ is the estimate $\hat{Y}_E$ of the population total Y computed from the sample drawn in the simulation $b \in \{1,...,B\}$, and B is the number of runs.

The same summary measures for Monte Carlo Simulations where the finite population $\mathcal{F}$ is recreated from a subpopulation for each simulation run drawn is

$$RB\left(\hat{Y}_E\right)\% = 100 \times \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{Y}_{E,b} - Y_b}{Y_b}, \tag{A.32}$$

$$MSE\left(\hat{Y}_E\right) = \frac{\sum_{b=1}^{B}\left(\hat{Y}_{E,b} - Y_b\right)^2}{B}, \tag{A.33}$$

$$RRMSE = \sqrt{\frac{MSE\left(\hat{Y}_E\right)}{Y_b^2}}, \text{ and} \tag{A.34}$$

$$RE\left(\hat{Y}_E\right)\% = 100 \times \left(\frac{MSE\left(\hat{Y}_{HT}\right)}{MSE\left(\hat{Y}_E\right)} - 1 \right), \tag{A.35}$$

where $\hat{Y}_E$ is the estimator being evaluated and $\hat{Y}_{E,b}$ is the estimate $\hat{Y}_E$ of the population total Y computed from the sample drawn in the simulation $b \in \{1,...,B\}$, and B is the number of runs.

A.5 Derivation of the Linear PA Estimator

In this section, we derive the linear PA estimator or the PA estimator with the linear working model using matrix algebra (see Section 1.7.3 and Definition 1.22 for details of linear PA estimators).

Let y be the outcome variable with an assumed linear superpopulation model $\mathcal{M}_y$ with $y_k \mid \mathbf{x}_k \overset{iid}{\sim} \mathcal{N}\left(\mathbf{x}_k\boldsymbol{\beta},\sigma_0^2\right)$, where $\mathbf{x}_k = \left(x_1,...,x_P\right) \in \mathbb{R}^{1\times P}$ is the vector with P-auxiliary variables, and $\boldsymbol{\beta} = \left(\beta_1,...,\beta_P\right)^{\mathrm{T}} \in \mathbb{R}^{P\times 1}$ is the vector of the regression coefficients of the linear predictor of the location parameter of the model $\mathcal{M}_y$. Let $\mathcal{F} = \left(\mathbf{y},\mathbf{x}\right)$ be the generated finite population that is N *iid* realizations of $\mathcal{M}_y$. The population $\mathcal{F}$ is sampled according to a sample design $p\left(\mathbf{S}=\mathbf{s}\right)$ that meets the suitable regularity conditions described in Section 5.9. Let $\mathbf{S} \in \left\{0,1\right\}^{N\times 1}$ be the discrete random vector for the sample membership indicator defined by $\mathbb{E}\left(\mathbf{S} \mid \mathcal{F}\right) = \boldsymbol{\pi} \in \left(0,1\right)^{N\times 1}$ and $\mathbb{V}\left(\mathbf{S} \mid \mathcal{F}\right) = \boldsymbol{\Delta} \in \mathbb{R}^{N\times N}$.

We are interested in estimating the population total of y in $\mathcal{F}$, defined as $Y = \sum_{k\in U} y_k$, using the auxiliary variables $\mathbf{x}$ observed in the sample and the known population totals $\mathbf{X}$. To compute the PA estimator, we need to estimate $\hat{\mu}_{pa,k} = \hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\hat{\boldsymbol{\beta}}_{pmle}$, that is, we first need to compute the PMLs of regression coefficients $\boldsymbol{\beta}$ of the model $\mathcal{M}_y$, fitted to the sample as

$$\hat{\boldsymbol{\beta}}_{pmle} = \underset{\boldsymbol{\beta}\in\mathcal{M}_y}{\arg\max}\ \log\mathcal{L}\left(\boldsymbol{\beta},\sigma;\mathbf{S},\mathbf{d},\mathbf{x}\,|\,\mathcal{F}\right), \tag{A.36}$$

where the sample-based log-likelihood of this model $\mathcal{M}_y$ is (1.3). The pseudo-log-likelihood in matrix notation is

$$\begin{aligned}\log\mathcal{L}\left(\boldsymbol{\beta},\sigma;\mathbf{S},\mathbf{d},\mathbf{x}\,|\,\mathcal{F}\right) = &-\frac{1}{2\sigma^2}\left(\mathbf{S}\odot\mathbf{d}\odot\left(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\right)\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\right)\\ &-\mathbf{S}^{\mathrm{T}}\mathbf{d}\left(\log(\sigma)\frac{\log(2\pi)}{2}\right)\end{aligned}. \tag{A.37}$$

The score function, $\mathcal{S}\left(\boldsymbol{\beta}\,|\,\mathcal{F}\right)$, is the vector with the partial derivatives of the PLL with respect to $\boldsymbol{\beta}$ given by

$$\begin{aligned}\mathcal{S}\left(\boldsymbol{\beta}\,|\,\mathcal{F}\right) &= \frac{\partial\log\mathcal{L}\left(\boldsymbol{\beta}\,|\,\mathcal{F}\right)}{\partial\boldsymbol{\beta}}\\ &= -\frac{1}{2\sigma^2}\left\{\left(\mathbf{S}\odot\mathbf{d}\odot\left(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\right)\right)^{\mathrm{T}}\mathbf{x}+\left(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)\right\}\end{aligned}.$$

The PMLEs are the roots of the score function set to zero

$$\mathcal{S}\left(\boldsymbol{\beta}\,|\,\mathcal{F}\right) = \left(\mathbf{S}\odot\mathbf{d}\odot\left(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\right)\right)^{\mathrm{T}}\mathbf{x}+\left(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)=0.$$

Solving for $\boldsymbol{\beta}$, we obtain the following

$$\left(\mathbf{S}\odot\mathbf{d}\odot\left(\mathbf{y}-\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)^{\mathrm{T}}\mathbf{x}+\left(\mathbf{y}-\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}}\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)=0$$

$$\left(\mathbf{S}\odot\mathbf{d}\odot\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)^{\mathrm{T}}\mathbf{x}+\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}}\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)=\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{y}\right)^{\mathrm{T}}\mathbf{x}+\mathbf{y}^{\mathrm{T}}\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}$$

$$2\left(\mathbf{S}\odot\mathbf{d}\odot\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)^{\mathrm{T}}\mathbf{x}=2\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{y}\right)^{\mathrm{T}}\mathbf{x}$$

$$\mathbf{x}^{\mathrm{T}}\left(\mathbf{S}\odot\mathbf{d}\odot\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)=\mathbf{x}^{\mathrm{T}}\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{y}\right)$$

$$\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}=\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{y}$$

$$\Rightarrow\quad\hat{\boldsymbol{\beta}}_{pmle}=\left(\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{x}\right)^{-1}\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{y}=\hat{\mathbf{T}}_{\mathbf{xx}}^{-1}\hat{\mathbf{T}}_{\mathbf{xy}}\,.\qquad\text{(A.38)}$$

where $\hat{\mathbf{T}}_{\mathbf{xx}}=\sum_{k\in U}S_k d_k\mathbf{x}_k^{\mathrm{T}}\mathbf{x}_k$ and $\hat{\mathbf{T}}_{\mathbf{xy}}=\sum_{k\in U}S_k d_k\mathbf{x}_k^{\mathrm{T}}\mathbf{y}_k$ are the HT estimators of the population matrix $\mathbf{T}_{\mathbf{xx}}=\mathbf{x}^{\mathrm{T}}\mathbf{x}$ and population vector $\mathbf{T}_{\mathbf{xy}}=\mathbf{x}^{\mathrm{T}}\mathbf{y}$ with the cross sums of $\mathbf{x}$ and $\mathbf{y}$.

Replacing the PA adjusted fitted mean of the model, $\hat{\boldsymbol{\mu}}_{pa}=\left(\mathbf{S}\odot\mathbf{x}\right)\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\hat{\boldsymbol{\beta}}_{pmle}$ and using the sampling weight $\mathbf{w}=\mathbf{d}$ in the generic expression in (1.25), the PA estimator of the total Y using the linear working model $\mathcal{M}_y$ is

$$\begin{aligned}\hat{Y}_{PA}&=\mathbf{d}^{\mathrm{T}}\left(\hat{\boldsymbol{\mu}}_{pa}\odot\mathbf{S}\right)\\&=\mathbf{d}^{\mathrm{T}}\left(\left(\mathbf{x}\hat{\boldsymbol{\Gamma}}_{\mathbf{X}}\hat{\boldsymbol{\beta}}_{pmle}\right)\odot\mathbf{S}\right).\\&=\mathbf{X}\hat{\boldsymbol{\beta}}_{pmle}\end{aligned}$$

which matches the expression in (1.37).

REMARK A.1. The derivation of the expression for $\hat{\boldsymbol{\beta}}_{pmle}$ in (A.38) is based on direct operations of Hadamard products. The expression of $\hat{\boldsymbol{\beta}}_{pmle}$ can be alternatively derived using rewriting the operation as a product of diagonal matrix and using the commutative property of the symmetric matrices.

The Hadamard product of the vector $\mathbf{S}$ and the matrix $\mathbf{A}$ is defined as

$$\mathbf{S}\odot\mathbf{A}=\begin{pmatrix} S_1 \\ S_2 \\ \dots \\ S_{N-1} \\ S_N \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N-1} & A_{1N} \\ A_{21} & A_{21} & \dots & A_{2N-1} & A_{2N} \\ \dots & \dots & \ddots & & \\ A_{N-11} & A_{N-11} & \dots & A_{N-11} & A_{N-1N} \\ A_{N1} & A_{N2} & \dots & A_{NN-1} & A_{NN} \end{pmatrix}$$

$$=\begin{pmatrix} S_1A_{11} & S_1A_{12} & \dots & S_1A_{1N-1} & S_1A_{1N} \\ S_2A_{21} & S_2A_{21} & \dots & S_2A_{2N-1} & S_2A_{2N} \\ \dots & \dots & \ddots & & \\ S_{N-1}A_{N-11} & S_{N-1}A_{N-11} & \dots & S_{N-1}A_{N-11} & S_{N-1}A_{N-1N} \\ S_NA_{N1} & S_NA_{N2} & \dots & S_NA_{NN-1} & S_NA_{NN} \end{pmatrix}=\mathbf{A}\odot\mathbf{S}.$$

We can rewrite the Hadamard product as

$$\mathbf{S}\odot\mathbf{A}=diag(\mathbf{S})\mathbf{A}=\mathbf{D_S}\mathbf{A},$$

where $\mathbf{D_S}=diag(\mathbf{S})$ is the diagonal matrix of $\mathbf{S}$ defined as

$$\mathbf{D_S}=diag(\mathbf{S})=\mathbf{S}=\begin{pmatrix} S_1 & 0 & \dots & 0 & 0 \\ 0 & S_2 & \dots & 0 & 0 \\ \dots & \dots & \ddots & 0 & 0 \\ 0 & 0 & \dots & S_{N-1} & 0 \\ 0 & 0 & \dots & 0 & S_N \end{pmatrix}.$$

Since $\mathbf{D_S}$ is a symmetric matrix, then the following identities hold:

$$\left(\mathbf{D_S}\right)^{\mathrm{T}} = \mathbf{D_S} \text{ symmetric matrix},$$
$$\mathbf{D_S A} = \mathbf{A D_S} \text{ commutative property},$$
$$\mathbf{S} \odot \mathbf{d} \odot \mathbf{A} = diag\left(\mathbf{S}\right)\left(\mathbf{d} \odot \mathbf{A}\right) = diag\left(\mathbf{S}\right) diag\left(\mathbf{d}\right)\left(\mathbf{A}\right) = \mathbf{D_S D_d A} \ ,$$
$$\mathbf{D_S D_d} = diag(\mathbf{S} \odot \mathbf{d}) = \mathbf{D_{S \odot d}}, \text{ and}$$
$$\left(\mathbf{D_S D_d}\right)^{\mathrm{T}} = \left(\mathbf{D_{S \odot d}}\right)^{\mathrm{T}} = \mathbf{D_{S \odot d}} \text{ symmetric matrix}.$$

Then

$$\begin{aligned}
\mathbf{S} \odot \mathbf{d} \odot \mathbf{A} &= \mathbf{D_S D_d A} = \mathbf{A D_S D_d} = \mathbf{A} \odot \mathbf{S} \odot \mathbf{d} \text{ commutative properties} \\
&= \mathbf{D_S D_d A} = \mathbf{A D_d D_S} = \mathbf{A} \odot \mathbf{d} \odot \mathbf{S} \\
&= \mathbf{D_S D_d A} = \mathbf{D_S A D_d} = \mathbf{S} \odot \mathbf{A} \odot \mathbf{d} \\
&= \mathbf{D_S D_d A} = \mathbf{D_S A D_d} = \mathbf{S} \odot \mathbf{A} \odot \mathbf{d} \\
&= \mathbf{D_S D_d A} = \mathbf{D_S D_d A} = \mathbf{S} \odot \mathbf{d} \odot \mathbf{A} \\
&= \mathbf{D_S D_d A} = \mathbf{D_d D_S A} = \mathbf{d} \odot \mathbf{S} \odot \mathbf{A}
\end{aligned} .$$

Representing the Hadamard product as a matrix product of a diagonal matrix then we can solve for $\hat{\boldsymbol{\beta}}_{pmle}$ as the roots of score function as follows:

$$\left(\mathbf{D_{S \odot d}}\left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)^{\mathrm{T}} \mathbf{x} + \left(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}} \mathbf{D_{S \odot d}} \mathbf{x} = 0$$
$$\left(\mathbf{D_{S \odot d}}\mathbf{y} - \mathbf{D_{S \odot d}}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)^{\mathrm{T}} \mathbf{x} + \left(\mathbf{y}^{\mathrm{T}} - \left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}}\right)\mathbf{D_{S \odot d}} \mathbf{x} = 0 .$$
$$\left(\left(\mathbf{D_{S \odot d}}\mathbf{y}\right)^{\mathrm{T}} - \left(\mathbf{D_{S \odot d}}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)^{\mathrm{T}}\right)\mathbf{x} + \mathbf{y}^{\mathrm{T}}\mathbf{D_{S \odot d}}\mathbf{x} - \left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}} \mathbf{D_{S \odot d}} \mathbf{x} = 0$$
$$\left(\mathbf{D_{S \odot d}}\mathbf{y}\right)^{\mathrm{T}} \mathbf{x} - \left(\mathbf{D_{S \odot d}}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)^{\mathrm{T}} \mathbf{x} + \mathbf{y}^{\mathrm{T}}\mathbf{D_{S \odot d}}\mathbf{x} - \left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}} \mathbf{D_{S \odot d}} \mathbf{x} = 0$$

Then

$$\left(\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)\right)^{\mathrm{T}}\mathbf{x}+\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}}\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\mathbf{x}=\left(\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\mathbf{y}\right)^{\mathrm{T}}\mathbf{x}+\mathbf{y}^{\mathrm{T}}\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\mathbf{x}$$

$$\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}}\mathbf{x}+\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}}\mathbf{x}=\mathbf{y}^{\mathrm{T}}\left(\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\right)\mathbf{x}+\mathbf{y}^{\mathrm{T}}\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\mathbf{x}$$

$$2\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)^{\mathrm{T}}\mathbf{x}=2\mathbf{y}^{\mathrm{T}}\left(\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\right)\mathbf{x}$$

$$\mathbf{x}^{\mathrm{T}}\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\left(\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}\right)=\left(\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\right)\mathbf{x}^{\mathrm{T}}\mathbf{y}$$

$$\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\mathbf{x}^{\mathrm{T}}\mathbf{x}\hat{\boldsymbol{\beta}}_{pmle}=\left(\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\right)\mathbf{x}^{\mathrm{T}}\mathbf{y}$$

$$\Rightarrow\hat{\boldsymbol{\beta}}_{pmle}=\left(\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\mathbf{x}^{\mathrm{T}}\mathbf{x}\right)^{-1}\left(\mathbf{D}_{\mathbf{S}\odot\mathbf{d}}\right)\mathbf{x}^{\mathrm{T}}\mathbf{y}=\left(\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{x}\right)^{-1}\left(\mathbf{S}\odot\mathbf{d}\odot\mathbf{x}\right)^{\mathrm{T}}\mathbf{y}\,.$$

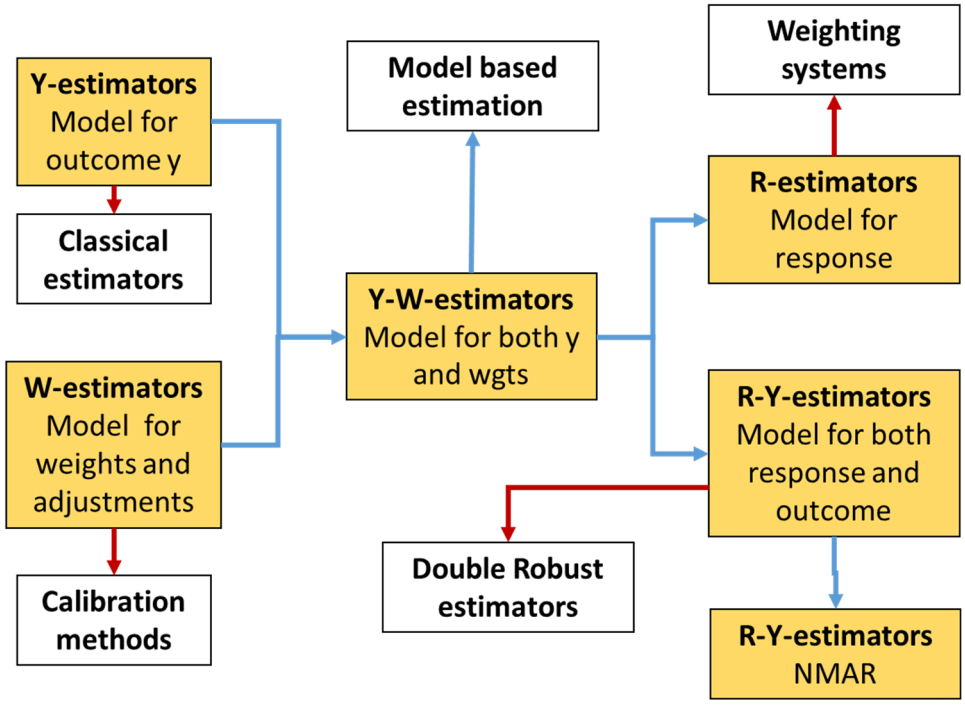
Appendix B Expanding the PA Approach

The approach presented in this dissertation attempts to unify estimation I survey theory by using a systematic approach based on standard statistical tools. The goal is to provide tools for answering current problems in estimation, in particular, estimation with nonresponse. Figure B.1 shows the areas of expansion of the PA framework. We classify these areas by the type of estimators shown below:

Estimator	Description
Y-estimators	Estimators of the outcome variable produced by replacing y in the estimator. The estimators presented in this dissertation are Y-estimators since the estimator is formed by using the fitted adjusted PLME means $\hat{\mu}_{pa,k}$.
W-estimators	Future development. Estimators of the outcome variable produced by replacing the sampling weight d by the fitted means of the distribution of an assumed model for the weights. The weights (or probabilities of inclusion) are assumed to be generated by a superpopulation model. These estimators establish a link from the PA approach to calibration and other methods for weighting adjustments.

Y-W-estimators	Future development. Combination of Y and W estimators, where the outcome and weights are replaced.
R-estimators	Future development. Estimators of the outcome variable produced by replacing the sampling weight d by fitted means of the distribution of an assumed model for the weights reflecting the effect nonresponse. The nonresponse adjusted weights are for the development of systems of weights for multipurpose surveys.
Y-R-estimators	Future development. Estimators of the outcome variable produced by replacing the sampling weight d and outcome variable reflecting the effect nonresponse.

Figure B.1 Future development areas of the PA framework



References

Agarwal, M., & Jain, N. (1989). A new predictive product estimator. *Biometrika, 76*, 822-823.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics, 16*(1), 3-14. doi:10.1016/0304-4076(81)90071-3

Arfken, G. B., Weber, H. J., & Harris, F. E. (2015). *Mathematical Methods for Physicists* (7th ed.). Waltham, MA: Elsevier.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika, 95*(3), 539-553. doi:10.1093/biomet/asn028

Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison. In P. Lahiri (Ed.), *Model selection* (pp. 135-207). Beachwood, OH: Institute of Mathematical Statistics. doi:10.1214/lnms/1215540968

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*(3), 279-292.

Binder, D. A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology, 22*(1), 17-22.

Binder, D., & Roberts, G. (2009). Design- and model-based inference for model parameters. In *Sample Surveys: Inference and Analysis Vol 29B.*

Breidt, F. J., & Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics, 28*(4), 1026-1053.

Breidt, F. J., & Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science, 32*(2), 190-205.

Breidt, F. J., Opsomer, J. D., & Sanchez-Borrego, I. (2016). Nonparametric variance estimation under fine stratification: an alternative to collapsed strata. *Journal of the American Statistical Association, 111*(514), 822-833. doi:10.1080/01621459.2015.1058264

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: a critical review. *Journal of Official Statistics, 29*, pp. 329-353.

Brick, J. M., Flores Cervantes, I., Lee, S., & Norman, G. (2011). Nonsampling errors in dual frame telephone surveys. *Survey Methodology, 37*(1), 1-12.

Burnham, K. P., & Anderson, D. R. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer.

Cardot, H., Degras, D., & Josserand, E. (2013). Confidence bands for Horvitz–Thompson estimators using sampled noisy functional data. *Bernoulli, 19*(5A), 3067-2097. doi:10.3150/12-BEJ443

Casella, G., & Berger, R. L. (2002). *Statistical inference.* Pacific Grove, CA: Duxbury Press.

Cassady, R., & Valiant, R. (1993). Conditional properties of poststratified estimators under normal theory. *Survey Methodology, 18*(2), 183-192.

Cassel, C., Särndal, C., & Wretman, J. (1977). *Foundations of inference in survey sampling.* New York, NY: Wiley.

Chambers, R. L., & Skinner, C. J. (1999). Intelligent calibration? *Proceedings of the Meeting of the International Association of Survey Statisticians*, (p. 221 231). Helsinki.

Chambers, R. L., Steel, D. G., Wang, S., & Welsh, A. (2012). *Maximum Likelihood Estimation for Sample Surveys* (1st ed.). Boca Raton, FL: Chapman and Hall/CRC.

Chang, C.-S. (2018). *Understanding Conditional Expectation via Vector Projection.* Retrieved from https://www.ee.nthu.edu.tw/cschang/Talk01142008.pdf

Chen, J. K., Valliant, R., & Elliott, M. R. (2018). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1-25. doi:10.1111/rssc.12327

Chen, J., & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica, 9*, 385-406.

Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J., . . . Thompson, M. (2017). Approaches to improving survey-weighted estimates. *Statistical Science, 32*(2), 227-248.

Cheng, R. (2017). *Non-standard parametric statistical inference.* Oxford: Oxford University Press.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2017). Double/Debiased Machine Learning for Treatment and Causal Parameters. Retrieved 02 13, 2018, from arXiv:1608.00060

Clifford, A. A. (1973). *Multivariate error analysis: A handbook of error propagation and calculation in many-parameter systems.* New York, NY: John Wiley & Sons.

Cochran, W. (1977). *Sampling Techniques* (3rd ed.). New Delhi: Wiley & Sons.

Cornfield, J. (1944). On samples from finite populations. *Journal of the American Statistical Association, 39*(226), 236-239. Retrieved from https://www.jstor.org/stable/2279953

Cox, D. R. (1970). *Simple Regression: Analysis of Binary Data.* London: Methuen Young Books.

Demnati, A., & Rao, J. (2004). Linearization variance estimation for survey data. *Survey Methodology, 30*(1), 17-26.

Deville, J., & Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*, 376-382.

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association, 88*(423), 1013-1020.

Dol, W., Steerneman, T., & Wansbeek, T. (1996). Matrix algebra and sampling theory: the case of the Horvitz-Thompson estimator. *Linear Algebra and its Applications, 237/238*, 225-238.

Efron, B., & Gous, A. (2001). Scales of evidence for model selection: Fisher versus Jeffreys. In P. Lahiri (Ed.), *Model selection* (pp. 208-246). Beachwood, OH: Institute of Mathematical Statistics. doi:10.1214/lnms/1215540972

Erdös, P., & Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Magyar Tuidoaniyos Akadenmia Budapest Matematikai Kutato Intezet Koezlemenyei, 4*, 49-57.

Fabrizi, E., & Lahiri, P. (2013). A design-based approximation to the Bayes Information Criterion in finite population sampling. *Statistica, LXXIII*(3), 289-301. doi:10.6092/issn.1973-2201/4325

Ferrar, S. L., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799-815.

Fuller, W. A. (1975). Regression analysis for sample surveys. *Sankhya C., 37*, 117-132.

Fuller, W. A. (2009). *Sampling Statistics.* Hoboken, New Jersey: John Wiley & Sons.

Fuller, W. A., & Isaki, C. T. (1981). Design under superpopulation models. In D. Krewski, R. Plateck, J. N. Rao, & M. P. Singh, *Current topics in survey sampling* (pp. 196-226). New York: Academic Press.

Ghanem, R., & Spanos, P. (2012). *Stochastic finite elements: a spectral approach.* Dover Publications; Revised edition.

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica, 52*(3), 681-700. doi:10.2307/1913471

Greene, W. H. (2008). *Econometric Analysis.* Upper Saddle River, NJ: Pearson Prentice Hall.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 1157-1182.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tuidoaniyos Akadenmia Budapest Matematikai Kutato Intezet Koezlemenyei*, 361-374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics, 35*(4), 1491-1523. doi:10.1214/aoms/1177700375

Hájek, J. (1971). Comment on an essay on the logical foundations of survey sampling by Basu, D. In V. Godambe, & D. e. Sprott, *Foundations of Statistical Inference* (p. 236). Holt McDougal.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory.* New York: John Wiley and Sons.

Hartley, H. O., & Ross, A. (1954). Unbiased ratio estimators. *Nature, 174*, 270-271.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd, corrected 12th printing 01/13/2017 ed.). New York: Springer-Verlag. doi:10.1007/978-0-387-84858-7

Haziza, D., & Beaumont, J.-F. (2017). Construction of weights in surveys: a review. *Statistical Science, 32*(2), 206-226. Retrieved from https://projecteuclid.org/euclid.ss/1494489812

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal. Biometrische Zeitschrift, 60*(3), 431-449. doi:10.1002/bimj.201700067

Horn, R., & Johnson, C. (2013). *Matrix Analysis* (2nd ed.). New York, NY: Cambridge University Press.

Horvitz, D., & Thompson, D. (1952). A generalisation of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663-685.

Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association, 77*(377), 89-96.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous Univariate Distributions* (2nd ed., Vol. I). Wiley.

Kim, J. K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica, 19*, 145-157.

Kim, J. K. (2010). Calibration estimation using exponential tilting in sample surveys. *Survey Methodology, 36*(2), 145-155.

Kim, J. K., & Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica, 24*(1), 375-394. doi:10.5705/ss.2012.005

Kim, J. K., & Riddles, M. K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology, 38*(2), 157-165.

Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: a recent survey. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 47-58.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology, 32*(2), 133-142.

Kott, P. S. (2016). Calibration weighting in survey sampling. *WIREs Computational Statistics, 8*, 39-53. doi:10.1002/wics.1374

Kott, P. S., & Liao, D. (2017). Calibration weighting for nonresponse that is not missing at random: allowing more calibration than response-model variables. *Journal of Survey Statistics and Methodology, 5*(2), 159-174. Retrieved from https://doi.org/10.1093/jssam/smx003

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* New York, NY, USA: Springer-Verlag.

Lehmann, E. (1999). *Elements of Large-Sample Theory.* New York, NY: Springer-Verlag.

Lehtonen, R., & Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology, 24*(1), 51-55.

Little, R. J. (2008). Weighting and prediction in sample surveys. *Calcutta Statistical Association Bulletin, 60*, 239-240.

Lohr, S. (2010). *Sampling: Design and Analysis* (2nd ed.). Boston: Brooks/Cole.

Luenberger, D. (1969). *Optimization by Vector Space Methods*. New York: John Wiley and Sons, Inc.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* doi:10.1002/9780470580066

Lumley, T. (2012). survey: analysis of complex survey samples. R package version 3.28-2.

Lumley, T., & Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology, 3*(1), 1-18.

Lumley, T., Shaw, P. A., & Dai, J. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data.

International Statistical Review, 79(2), 200-220. doi:10.1111/j.1751-5823.2011.00138.x

Madow, W. G. (1948). On the limiting distributions of estimates based on samples from finite universes. *The Annals of Mathematical Statistics, 19,* 535- 545.

Magnus, J. R., & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Chichester: Wiley.

McConville, K. S., Breidt, F. J., Lee, T. C., & Moisen, G. G. (2017). Model-assisted survey regression estimation with the Lasso. *Journal of Survey Statistics and Methodology, 5*(2), 131-158. doi:10.1093/jssam/smw041

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). New York, NY: Chapman and Hall/CRC.

Montanari, G. E. (1987). Post-sampling efficient Q-R prediction in large sample surveys. *International Statistical Review / Revue Internationale de Statistique, 50*(22), 191-202.

Montanari, G. E. (1998). On regression estimation of finite population means. *Survey Methodology, 24*(1), 69-77.

Montanari, G. E. (2002). Theory & methods: conditioning on auxiliary variable means in finite population inference. *Australian & New Zealand Journal of Statistics,* 407-421. doi:10.1111/1467-842X.00138

Montanari, G. E., & Ranalli, M. G. (2002). Asymptotically efficient generalized regression estimators. *Journal of Official Statistics, 18*(4), 577-589.

Montanari, G. E., & Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association, 100,* 1429-1442.

Mukhopadhyay, P. (2016). *Complex Surveys: Analysis of Categorical Data.* Singapore: Springer.

Murthy, M. (1964). Product method of estimation. *Sankhyā: The Indian Journal of Statistics, Series A, 26*(1), 69-74.

Nascimiento Silva, P., & Skinner, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology, 23*(1), 23-32.

Opsomer, J. D., Breidt, F. J., Moisen, G. G., & Kauermann, G. (2007). Rejoinder to Opsomer, Breidt, Moisen, and Kauermann (2007). *Journal of the American Statistical Association, 102*(478), 415-416.

Papke, L., & Wooldridge, J. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 619-632. doi:10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1

Pfeffermann, D., & Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya: The Indian Journal of Statistics, Series B, 61*(1), 166-186. Retrieved from http://www.jstor.org/stable/25053074

Polansky, A. (2011). *Introduction to statistical limit theory*. Boca Raton, FL: Chapman & Hall/CRC.

Posthuma Partners. (2018). lmvar: Linear Regression with Non-Constant Variances. Retrieved from https://CRAN.R-project.org/package=lmvar

R Development Core Team. (2017). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org

Rao, C. R., & Wu, Y. (2001). On model selection. In P. Lahiri (Ed.), *Model selection* (Vol. 38, pp. 1-57). Beachwood, OH: Institute of Mathematical Statistics. doi:10.1214/lnms/1215540960

Rao, J. N. K. (1966). Alternative estimators in pps sampling for multiple characteristics. *Sankhya: The Indian Journal of Statistics, Series A, 28*(1), 47-60. Retrieved from http://www.jstor.org/stable/25049398

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics, 10*(2), 153-165.

Rao, J. N. K. (2008). Discussion of "Weighting and prediction in sample surveys" by Little, R. *Calcutta Statistical Association Bulletin, 60*, 29-39.

Rao, P. (1971). Some notes on misspecification in multiple regressions. *The American Statistician, 25*(5), 37-39. Retrieved from https://www.jstor.org/stable/2686082

Rawlings, J. O., Pantula, S. G., & Dickey, A. D. (1998). *Applied regression analysis: A research tool* (2nd ed.). New York, NY: Springer-Verlag.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Journal of Applied Statistics, 54*(3), 507-554.

Rubin-Bleuer, S., & Schiopu Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics, 33*(6), 2789-2810.

Ruppert, D. (2007). Comment to Opsomer, Breidt, Moisen, and Kauermann (2007). *Journal of the American Statistical Association, 102*(478), 409-411.

Särndal, C., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* Chichester, England: John Wiley & Sons.

Särndal, C., Swensson, B., & Wretman, J. (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:10.2307/2336118

Särndal, C., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology, 33*(2), 99-119.

Scott, A., & Wu, C.-F. (1981). On the Asymptotic Distribution of Ratio and Regression Estimators. *Journal of the American Statistical Association, 7*(76), 98-102.

Shah, B. V. (2004). Comment to Demnati and Rao (2004): Linearization Variance Estimators for Survey Data. *Survey Methodology, 30*(1), 17-26.

Small, C. G. (2010). *Expansions and Asymptotics for Statistics.* Chapman & Hall.

Somol, P., Novovicova, J., & Pudil, P. (2010). Efficient feature subset selection and subset size optimization. In A. Herout (Ed.), *Pattern Recognition.* Rijeka: IntechOpen. doi:10.5772/9356

Srivastava, S. K., & Jhajj, H. S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika, 68*(1), 341-343. doi:10.1093/biomet/68.1.341

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R.* Boca Raton, FL: Chapman and Hall/CRC.

Stasinopoulos, M., Rigby, B., Voudouris, V., Akantziliotou, C., Enea, M., & Kiose, D. (2017). gamlss: Generalised Additive Models for Location Scale and Shape Version 5.1-2.

Sverchkov, M. (2010). On modeling and estimation of response probabilities when missing data are not missing at random . *Joint Statistical Meetings, Proceedings of the Survey Research Methods of the American Statistical Association.*

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematic Sciences),* 12–18.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: a review. In C. C. Aggarwal, *Data Classification: Algorithms and Applications* (pp. 37-64). Chapman and Hall/CRC.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B, 58,* 267-288.

Tillé, Y. (1999). Estimation in surveys using conditional inclusion probabilities: complex design. *Survey Methodology, 25*(1), 57-66.

Tillé, Y. (2006). *Sampling Algorithms.* Springer.

Tillé, Y., & Matei, A. (2016). sampling: Survey Sampling. *{R package version 2.8*. Retrieved from https://CRAN.R-project.org/package=sampling

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples.* New York: Springer.

Valliant, R., Dever, J. A., & Kreuter, F. (2018). PracTools: Tools for Designing and Weighting Survey Samples. *R package version 0.8*. Retrieved from https://CRAN.R-project.org/package=PracTools

Valliant, R., Dorfman, A., & Royall, R. (2000). *Finite population sampling and inference: a prediction approach.* New York: John Wiley & Sons.

van der Laan, M. J., & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data.* New York: Springer. doi:10.1007/978-1-4419-9782-1

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika, 61*(3), 439-447. Retrieved from www.jstor.org/stable/2334725

Williams, D. (2011). *Probability with Martingales* (14th ed.). Cambridge: Cambridge University Press.

Wolter, K. (2017). *Introduction to variance estimation* (2nd ed.). New York: Springer-Verlag.

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.

Woodruff, R., & Causey, B. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association, 71*(354), 315-321. doi:10.2307/2285303

Wu, C., & Sitter, R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association, 96*(453), 185-193.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R.* New York, NY, USA: Springer.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association (Theory and Methods), 101*(406), 1418-1429. doi:10.1198/016214506000000735