On high-dimensional variance estimation in survey sampling

Esther Eustache[(a)] , Mehdi, Dagdoug[(b)] and David Haziza[(c)]

(a) University of Neuchâtel, Institute of statistics,
Neuchâtel, Switzerland
(b) McGill University, Department of Mathematics and Statistics,
Montreal, Canada
(c) University of Ottawa, Department of Mathematics and Statistics,
Ottawa, Canada

Abstract

Utilizing predictive modeling at different survey stages can improve the accuracy of a point estimator or help tackle issues such as missing values. So far, the existing literature on predictive models for survey data has predominantly concentrated on scenarios with low-dimensional data, wherein the number of variables is small compared to the sample size. In this paper, assuming a linear regression model, we show that customary variance estimators based on a first Taylor expansion or jackknife may suffer from substantial bias in a high-dimensional setting. We explain why this is so through a mix of theoretical and empirical investigations. We propose some bias-adjusted variance estimators and show empirically that the proposed variance estimators perform well in terms of bias, even in a high-dimensional setting.

Key words: Bias-adjusted variance estimator; Generalized regression estimator; Jackknife variance estimation; Linear regression imputation; Taylor-based variance estimator.

1 Introduction

Predictive modeling can be applied at various stages of a survey to enhance the precision of a point estimator and to address the problem of missing values, among others. Using predictive models enables us to exploit a relationship between a survey variable Y and a

set of predictors $X_1, X_2, \ldots, X_p$. For instance, model-assisted estimation procedures use a set of predicted values to improve the efficiency of point estimators; e.g., see Särndal (1992) and Breidt and Opsomer (2017). To mitigate the potential nonresponse bias caused by item nonresponse, it is common practice to employ some form of imputation, which involves generating a set of predictions to substitute for the missing values; e.g., see Haziza (2009) and Chen and Haziza (2019).

The literature on predictive models for survey data has primarily focused on low-dimensional data settings, where the number of variables p is small relative to the sample size n. Formalized mathematically, it means that $p/n \to 0$. Some notable exceptions include Cardot et al. (2017), Ta et al. (2020), Chauvet and Goga (2022) and Dagdoug et al. (2022). With the advent of big data sets, moderate to high-dimensional settings are becoming more prevalent. In this article, a high-dimensional setting refers to a situation where the number of predictors p is of the same magnitude as the sample size n so that $p/n \to \mathcal{K} \in (0, 1)$.

High-dimensional linear regression models pose some challenges compared to traditional linear regression models with fewer predictors. In particular, the common variance estimation procedures tend to break down when $p/n \to \mathcal{K} \in (0, 1)$. Indeed, variance estimators based on a first-order Taylor expansion procedure tend to underestimate the variance of point estimators, whereas resampling procedures such as the jackknife and the bootstrap may lead to substantial overestimation of the true variance; see Wolter (2007) and Mashreghi et al. (2016) for a review of resampling methods in finite population sampling.

The contributions of this paper are: (i) we explain why variance estimators based on a first-order Taylor and jackknife variance estimators tend to breakdown through a mix of empirical and theoretical investigations. We consider two different setups that involve the customary linear regression model: (a) The model-assisted estimation setup through the use of the generalized regression estimator (see, e.g., Särndal, 1980; Särndal, 1992; Särndal, 2007); (b) The

deterministic linear regression imputation setup (e.g., Chen and Haziza, 2019). (ii) In the context of Bernoulli sampling and simple random sampling without replacement, we propose some bias-adjusted variance estimators that are shown to work well, at least in our experiments.

We adopt the following notations. Let $U := \{1, 2, ..., N\}$ be a finite population of size N. Our interest lies in estimating the finite population mean

$$\mu_y := \frac{1}{N} \sum_{k \in U} y_k,$$

of a survey variable Y, where y_k denotes the y-value attached to unit k. We select a sample, S, of (expected) size n, according to a probability sampling design $\mathcal{P}(S \mid \boldsymbol{Z})$, where $\boldsymbol{Z} \in \mathbb{R}^{N \times d}$ denotes the matrix of design information. We restrict our attention to non-informative sampling design; see, e.g., Pfeffermann and Sverchkov (2009). The sample S is fully characterized by the vector of sample selection indicators, $\boldsymbol{I} := [I_1, I_2, \ldots, I_N]^\top$, where $I_k := 1$ if $k \in S$, and $I_k := 0$, otherwise. We denote by $\pi_k := \mathbb{P}(I_k = 1) > 0$ and $\pi_{k\ell} := \mathbb{P}(I_k = 1, I_\ell = 1) > 0$, for $k, \ell \in U$, the first-order and the second-order inclusion probabilities, respectively.

2 Linear prediction in survey sampling

We consider the customary linear regression model:

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \epsilon_k, \qquad k \in U, \tag{1}$$

where $\boldsymbol{\beta}$ is a p-vector of unknown coefficients and the errors ϵ_k satisfy $\mathbb{E}[\epsilon_k | \mathbf{x}_k] = 0$, $\mathbb{E}\left[\epsilon_k^2 | \mathbf{x}_k\right] := \sigma^2 < \infty$ and are independently and identically distributed. We assume that the intercept is included in the covariates; i.e., the first component of $\mathbf{x}_k$ is 1 for all $k \in U$. Although we assume an homoscedastic variance structure, our results can be easily extended to the case of an heteroscedastic variance structure.

Below, we use the notation $\boldsymbol{y}_S \in \mathbb{R}^{n_s}$ and $\boldsymbol{X}_S \in \mathbb{R}^{n_s \times p}$ to denote the vector of y-values and the design matrix corresponding to the sample, respectively. Also, we use $\boldsymbol{\Pi}_S \in \mathbb{R}^{n_s \times n_s}$ to

denote the diagonal matrix, whose kth diagonal element is π_k.

2.1 Model-assisted estimation

In this section, we assume that the observed data are given by

$$\mathcal{D}_{ma} := \{(\mathbf{x}_k, y_k) \ ; \ k \in S\}.$$

In addition, we assume that the vector of population totals,

$$\mathbf{t}_{\mathbf{x}} := \left[\sum_{k\in U} x_{k1}, \sum_{k\in U} x_{k2}, ..., \sum_{k\in U} x_{kp}\right]^\top,$$

is available from an external source. The Generalized REGression (GREG) estimator of μ_y is given by

$$\widehat{\mu}_{greg} := \frac{1}{N}\left(\sum_{k\in U} \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_S + \sum_{k\in S} \frac{y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_S}{\pi_k}\right), \tag{2}$$

where

$$\widehat{\boldsymbol{\beta}}_S = \left(\boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S \boldsymbol{\Pi}_S^{-1} \boldsymbol{y}_S$$

is defined as the weighted least squares estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_S := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{k\in S} \frac{\left(y_k - \mathbf{x}_k^\top \boldsymbol{\beta}\right)^2}{\pi_k}. \tag{3}$$

Throughout the paper, we assume that the $p \times p$ matrix, $\boldsymbol{A}_{\Pi S} := \boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{X}_S$, is non-singular.

In our setting, the GREG estimator can be written in the so-called projection form:

$$\widehat{\mu}_{greg} = \frac{1}{N} \sum_{k\in U} \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_S$$

since

$$\sum_{k\in S} \pi_k^{-1} \widehat{\epsilon}_{kS} = 0,$$

where $\widehat{\epsilon}_{kS} := y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_S$ denotes the sample residual attached to unit $k \in S$, (Särndal et al., 1992, Chapter 6).

In this section, as the reference distribution for studying the properties of variance estimators, we use the joint distribution induced by the superpopulation model (1) and the sampling design. Consider the following decomposition:

$$\mathbb{V}_{mp}\left(\widehat{\mu}_{greg}\right) = \mathbb{E}_m\left[\mathbb{V}_p\left(\widehat{\mu}_{greg}\right)\right] + \mathbb{V}_m\left(\mathbb{E}_p\left[\widehat{\mu}_{greg}\right]\right), \tag{4}$$

where the subscripts p and m are used to denote the sampling design and the imputation model, respectively.

Based on (4), an estimator of the variance of $\widehat{\mu}_{greg}$ based on a first-order Taylor expansion is given by

$$\widehat{V}_{\text{tay}} = \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\epsilon}_{kS}}{\pi_k} \frac{\widehat{\epsilon}_{\ell S}}{\pi_\ell} + \frac{\widehat{\sigma}^2}{N}, \tag{5}$$

where $\widehat{\sigma}^2$ denotes an unbiased estimator of σ^2. The first term on the right hand-side of (5) is the focus in this article.

Särndal et al. (1989) advocated the use of a g-weighted version, which is obtained from (5) by replacing $\widehat{\epsilon}_{kS}$ with $g_k \times \widehat{\epsilon}_{kS}$, where

$$g_k := 1 + \left(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x},\pi}\right)^\top \boldsymbol{A}_{\Pi S}^{-1} \mathbf{x}_k = \mathbf{t}_{\mathbf{x}}^\top \boldsymbol{A}_{\Pi S}^{-1} \mathbf{x}_k, \qquad k \in S, \tag{6}$$

is the so-called g-weight attached to unit $k \in S$, with $\widehat{\mathbf{t}}_{\mathbf{x},\pi}$ denoting the Horvitz–Thompson estimator of $\mathbf{t}_{\mathbf{x}}$. The second equality in (6) is satisfied as the intercept is included in the set of predictors and the variance structure is assumed to be homoscedastic. This leads to

$$\widehat{V}_g = \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{g_k \widehat{\epsilon}_{kS}}{\pi_k} \frac{g_\ell \widehat{\epsilon}_{\ell S}}{\pi_\ell} + \frac{\widehat{\sigma}^2}{N}. \tag{7}$$

Jackknife variance estimation for the GREG estimator has been discussed in Yung and Rao (1996), Duchesne (2000) and Valliant (2002), among others. Here, we consider the generalized

jackknife variance estimator of Campbell (1980) and Berger and Skinner (2005). Let $\widetilde{h}_{kk}^{\pi} := \mathbf{x}_k^{\top} \boldsymbol{A}_{\Pi S}^{-1} d_k \mathbf{x}_k$ be the survey weighted leverage of element $k \in S$, with $d_k = \pi_k^{-1}$. The next proposition gives a closed-form expression of the generalized jackknife variance estimator in the case of the GREG estimator.

Proposition 2.1. *An estimator of (4) based on the generalized jackknife variance estimator of Berger and Skinner (2005) has a closed-formed formula given by*

$$\widehat{V}_{\text{jack}} = \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{(1 - w_k)\, g_k\, \widehat{\epsilon}_{kS}}{\left(1 - \widetilde{h}_{kk}^{\pi}\right) \pi_k} \frac{(1 - w_\ell)\, g_\ell\, \widehat{\epsilon}_{\ell S}}{\left(1 - \widetilde{h}_{\ell\ell}^{\pi}\right) \pi_\ell} + \frac{\widehat{\sigma}^2}{N}, \tag{8}$$

where $w_k := (N\pi_k)^{-1}$ *for* $k \in S$.

Proof. See Appendix A. ■

2.2 Deterministic linear regression imputation

Predictions based on a linear regression model are also used in the context of imputation for item nonresponse. In this context, the survey variable Y is observed only for a subset $S_r \subseteq S$, called the set of respondents to item Y. We denote by $S_m = S - S_r$ the set of nonrespondents to item Y. Let $\boldsymbol{R} := [R_1, R_2, \ldots, R_N]^{\top}$ be the N-vector of response indicators, where $R_k = 1$ if $k \in S_r$, and $R_k = 0$, otherwise. Here, the predictors $X_1, \ldots, X_p$, are assumed to be available for both the respondents and the nonrespondents. We assume that: (i) The data $\{(\mathbf{x}_k, y_k, r_k)\}_{k \in U}$ are identically and independently distributed; (ii) The data are Missing At Random (Rubin, 1976, MAR):

$$\mathbb{P}\left(R_k = 1 | \mathbf{x}_k, y_k\right) = \mathbb{P}\left(R_k = 1 | \mathbf{x}_k\right);$$

(iii) The positivity assumption is satisfied; i.e., $\mathbb{P}\left(R_k = 1 | \mathbf{x}_k\right) > 0$, almost surely. Available to the imputer are the data

$$\mathcal{D}_{imp} := \{(\mathbf{x}_k, y_k) \ ; \ k \in S_r\} \bigcup \{\mathbf{x}_k \ ; \ k \in S_m\}.$$

Imputation consists of estimating the relationship between Y and $X_1, \ldots, X_p$ based on the respondents and extrapolating this relationship to the set of nonrespondents.

An estimator of μ_y after deterministic linear imputation is given by

$$\widehat{\mu}_{lr} := \frac{1}{\widehat{N}} \left(\sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_R}{\pi_k} \right) = \frac{1}{\widehat{N}} \sum_{k \in S} \frac{\widetilde{y}_k}{\pi_k}, \tag{9}$$

where $\widehat{N} := \sum_{k \in S} \pi_k^{-1}$ and $\widetilde{y}_k := R_k y_k + (1 - R_k)\, \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_R$ with

$$\widehat{\boldsymbol{\beta}}_R = \left(\sum_{k \in S_r} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\pi_k} \right)^{-1} \sum_{k \in S_r} \frac{\mathbf{x}_k y_k}{\pi_k} = \left(\boldsymbol{X}_R \boldsymbol{\Pi}_R^{-1} \boldsymbol{X}_R^\top \right)^{-1} \boldsymbol{X}_R \boldsymbol{\Pi}_R^{-1} \boldsymbol{y}_R \tag{10}$$

is defined as the weighted least squares estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_R := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \sum_{k \in S_r} \frac{\left(y_k - \mathbf{x}_k^\top \boldsymbol{\beta} \right)^2}{\pi_k}. \tag{11}$$

In (10), the quantities $\boldsymbol{X}_R$, $\boldsymbol{\Pi}_R$ and $\boldsymbol{y}_R$ correspond to the counterparts of $\boldsymbol{X}_S$, $\boldsymbol{\Pi}_S$ and $\boldsymbol{y}_S$, respectively, restricted to the set of respondents S_r.

To estimate the variance of $\widehat{\mu}_{lr}$, we consider the reverse framework, originally proposed by Fay (1991) and Shao and Steel (1999); see also Kim and Rao (2009) and Haziza and Vallée (2020). Using this framework, the total variance of $\widehat{\mu}_{lr}$ can be expressed as

$$\mathbb{V}(\widehat{\mu}_{lr}) = \mathbb{E}_m \mathbb{E}_q \mathbb{V}_p(\widehat{\mu}_{lr}) + \mathbb{E}_q \mathbb{V}_m \mathbb{E}_p(\widehat{\mu}_{lr} - \mu_y), \tag{12}$$

where the subscript q denotes the nonresponse mechanism. Let us define $\widehat{h}_{k\ell}^{\pi} := \mathbf{x}_k^\top \boldsymbol{A}_{\Pi R}^{-1} d_k \mathbf{x}_\ell$ and $\widehat{\Gamma}_k := \sum_{\ell \in S_m} \widehat{h}_{k\ell}^{\pi}$, where $\boldsymbol{A}_{\Pi R} := \boldsymbol{X}_R^\top \boldsymbol{\Pi}_R^{-1} \boldsymbol{X}_R$. Again, we assume that the matrix $\boldsymbol{A}_{\Pi R}$ is non-singular.

Proposition 2.2. *An estimator of the variance of $\widehat{\mu}_{lr}$ based on a first-order Taylor expansion is given by*

$$\widehat{V}_{I,\text{tay}} = \frac{1}{\widehat{N}^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\xi}_k - \widehat{\mu}_{lr}}{\pi_k} \frac{\widehat{\xi}_\ell - \widehat{\mu}_{lr}}{\pi_\ell} + \frac{\sigma^2}{\widehat{N}^2} \sum_{k \in S_r} \frac{1}{\pi_k} \left\{ 1 - R_k (1 + \widehat{\Gamma}_k) \right\}^2, \tag{13}$$

where

$$\widehat{\xi}_k := \widetilde{y}_k + r_k \widehat{\Gamma}_k \, \widehat{\epsilon}_{kR} \tag{14}$$

with $\widehat{\epsilon}_{kR} = y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_R$, $k \in S_r$.

The proof of Proposition 2.2 is straightforward and is thus omitted.

We now turn to jackknife variance estimation in the context of deterministic linear regression. Berger and Rao (2006) extended the results of Berger and Skinner (2005) to the case of mean and ratio imputation. We extend the results of Berger and Rao (2006) to the more general setting of deterministic linear regression imputation. Next, we provide an estimator of the total variance in (12) based on the generalized jackknife of Berger and Rao (2006) in the context of deterministic linear regression imputation and exhibit a closed-form expression.

Result 2.1. *An estimator of the variance of* $\widehat{\mu}_{lr}$ *based on the generalized jackknife variance estimator of Berger and Rao (2006) has a closed-form expression given by*

$$\widehat{V}_{I,\text{jack}} = \frac{1}{\widehat{N}^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\mu}_{lr} - \widehat{\xi}_k^{(jk)}}{\pi_k} \frac{\widehat{\mu}_{lr} - \widehat{\xi}_\ell^{(jk)}}{\pi_\ell} + \frac{\sigma^2}{\widehat{N}^2} \sum_{k \in S_r} \frac{1}{\pi_k} \left\{1 - R_k(1 + \widehat{\Gamma}_k)\right\}^2, \tag{15}$$

where

$$\widehat{\xi}_k^{(jk)} := \widetilde{y}_k + r_k \widehat{\Gamma}_k \frac{\widehat{\epsilon}_{kR}}{1 - \widehat{h}_{kk}^{\pi}}. \tag{16}$$

Proof. See Appendix B. ■

3 Behavior of some commonly used variance estimators: Empirical studies

In this section, we present the results of two simulation studies. In Section 3.1, we first examine the empirical performance of the variance estimators described in Section 2.1, whereas Section 3.2 considers the variance estimators discussed in Section 2.2. In the model-assisted setup, we denote by $\mathcal{K} = p/n$, the ratio of the number of predictors to the expected sample size. In the linear regression imputation setup, we denote by $\mathcal{K} = p/\mathbb{E}[n_r]$, the ratio of the number of predictors to the expected number of respondents.

3.1 Model-assisted estimation: the GREG estimator

We generated a finite population U of size $N = 5,000$ consisting of 223 explanatory variables $X_1, \ldots, X_{223}$, and a survey variable Y. The variables $X_1, \ldots, X_{223}$, were generated from a multivariate normal distribution with a mean vector equal to $5 \times \mathbf{1}^\top$ and correlation matrix, whose diagonal elements were equal to 1 and off-diagonal elements equal to 0.3, where $\mathbf{1}$ denotes the vector of ones. Given the X-variables, we generated a survey variable Y according to the linear regression model

$$y_k = 14 - 4x_{1k} + 3x_{2k} + 4x_{3k} + \epsilon_k, \tag{17}$$

where the errors ϵ_k were generated from a normal distribution with mean equal to 0 and variance equal to 25^2. This led to a model R^2 approximately equal to 0.6. In (17), note that only the first three variables X_1, X_2, and X_3 were used for generating the Y-variable.

From the population, we selected $R = 10,000$ samples of (expected) sample size $n = 300$, according to two sampling designs: simple random sampling design without replacement and Bernoulli sampling. In each sample, we computed several GREG estimators, $\widehat{\mu}_{greg}$, given by (2), based on different sets of explanatory variables. In addition to X_1, X_2 and X_3, we included a number of noise variables denoted by p_{noise}. The values for p_{noise} were set to: 0, 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, and 220. This led to 12 estimators, $\widehat{\mu}_{greg}$, of μ_y. To estimate the variance of $\widehat{\mu}_{greg}$, we computed $\widehat{V}_{\text{tay}}$ given by (5), $\widehat{V}_g$ given by (7), and $\widehat{V}_{\text{jack}}$ given by (8).

As a measure of bias of a variance estimator, we computed its Monte Carlo percent relative bias (RB). Using the generic notations $\widehat{\mu}$ and $\widehat{V}$ for a point and a variance estimator, respectively, the RB of $\widehat{V}$ is defined as

$$RB(\widehat{V}) := 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{\widehat{V}^{(r)} - V_{MC}\left(\widehat{\mu}\right)}{V_{MC}\left(\widehat{\mu}\right)}, \tag{18}$$

where $V_{MC}\left(\widehat{\mu}\right)$ denotes the Monte-Carlo variance of $\widehat{\mu}$ and $\widehat{V}^{(r)}$ denotes the estimator $\widehat{V}$ at the rth iteration, $r = 1, \ldots, 10,000$. The results for simple random sampling without replacement and Bernoulli sampling are shown in Figures 1 and 2, respectively.

From Figures 1 and 2, we note that the three variance estimators performed well for small values of p/n. For instance, for $p/n = 3/300$, which corresponds to the case of $p_{noise} = 0$, the estimator $\widehat{V}_{\text{tay}}$ exhibited a value of RB of about 8.9% for Bernoulli sampling and 4.2% for simple random sampling without replacement. The g-weighted version $\widehat{V}_g$ showed a bias of 4.7% for Bernoulli sampling and of 5.2% for simple random sampling without replacement. The jackknife variance estimator $\widehat{V}_{\text{jack}}$ showed a bias of about 7.3% for Bernoulli sampling and 7.9% for simple random sampling without replacement. However, for $p/n = 83/300 \approx 0.28$, the RB of $\widehat{V}_{\text{tay}}$ was equal to -44% in the case of Bernoulli sampling and -43.2% in the case of simple random sampling without replacement. The magnitude of the underestimation got worse as p/n increased. The g-weighted version $\widehat{V}_g$ did better than $\widehat{V}_{\text{tay}}$ with a value of RB equal to -25.1% for Bernoulli sampling, and equal to -24.4% for simple random sampling without replacement. On the other hand, the jackknife variance estimator exhibited significant overestimation with values of RB equal to 41.5% in the case of Bernoulli sampling and 42.5% in the case of simple random sampling without replacement. The magnitude of the bias increased as $\mathcal{K}$ increased.

3.2 Deterministic linear regression imputation

We started by generating 5,000 realizations of a vector of explanatory variables, of size 113, from a multivariate normal distribution with a mean vector equal to $5 \times \mathbf{1}^\top$ and correlation matrix, whose diagonal elements were equal to 1 and the off-diagonal elements were equal to 0.3. We then repeated $R = 10,000$ iterations of the following process:

(i) Given the explanatory variables, we generated the survey variable Y according to Model

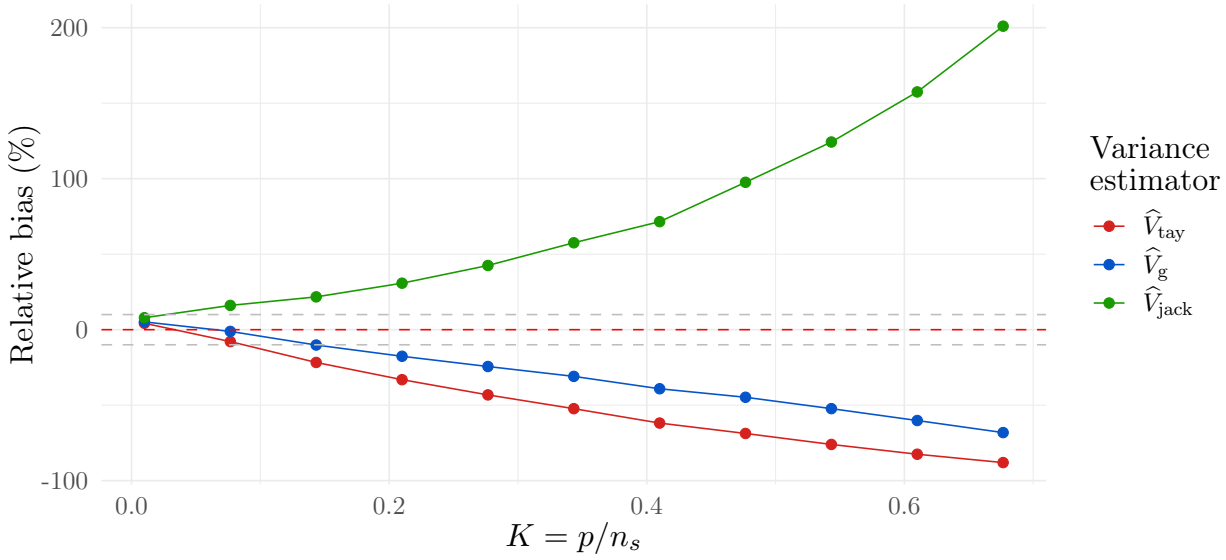

Figure 1: Behaviour of two variance estimators for $\widehat{\mu}_{greg}$ under simple random sampling without replacement.

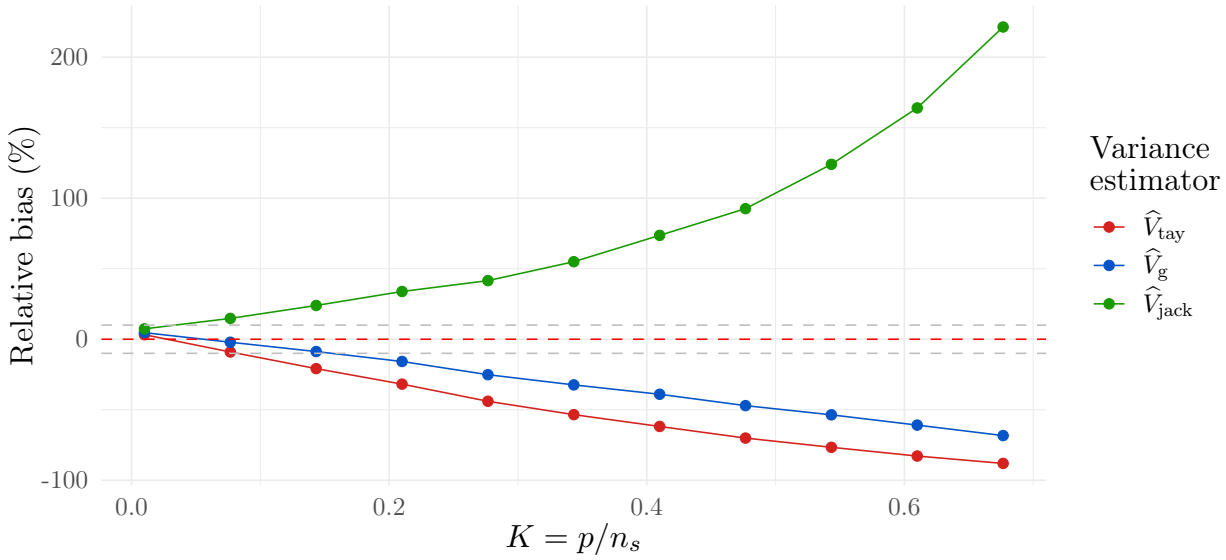

Figure 2: Behaviour of two variance estimators for $\widehat{\mu}_{greg}$ under Bernoulli sampling.

(17).

(ii) From the finite population of size $N = 5,000$ generated in Step (i), a sample, of (expected) size $n = 300$, was selected according to (1) simple random sampling without replacement and (2) Bernoulli sampling.

(iii) In each sample, the response indicators R_k, $k \in S$, were independently generated according to a Bernoulli distribution with probability

$$p_k = \{1 + \exp(1 + \lambda_1 x_{1k} + \lambda_2 x_{2k} + \lambda_3 x_{3k})\}^{-1},$$

where the values of λ_1-λ_3 were set to obtain an overall response rate of about 50%. Thus, in each sample, the expected number of respondents, $\mathbb{E}(n_r)$, was equal to 150.

(iv) The missing values in each sample were imputed through deterministic linear regression imputation with different subsets of explanatory variables. The first subset of explanatory variables included the variables X_1, X_2, and X_3 only, corresponding to the true model. In addition to X_1, X_2 and X_3, we included a number of noise variables denoted by p_{noise}. This led to 12 sets of explanatory variables of size p, where $p = p_{noise} + 3$. The values for p_{noise} were set to: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, and 110. As a result the ratio $\mathcal{K} = p/\mathbb{E}(n_r)$ ranged from 3/150 to 113/150. Each of the 12 models was fitted on the set of responding units, which led to 12 sets of imputed values.

(v) For each of the 12 sets of imputed values, we computed the imputed estimator $\widehat{\mu}_{lr}$ given by (9), leading to a set of 12 imputed estimators.

(vi) We estimated the variance of the 12 imputed estimators using two variance estimators: (i) The variance estimator based on a first-order Taylor expansion denoted by $\widehat{V}_{I,\text{tay}}$; see Section 2.2; and (ii) The generalized jackknife variance estimator, denoted by $\widehat{V}_{I,\text{jack}}$; see Section 2.2.

As a measure of relative bias of a variance estimator $\widehat{V}$, we computed its Monte Carlo percent relative bias (RB) given by (18).

From Figures 3 and 4, we note that both $\widehat{V}_{I,\text{tay}}$ and $\widehat{V}_{I,\text{jack}}$ performed well for small values of $\mathcal{K}$. For instance, for $\mathcal{K} = 3/150$, which corresponds to the case of $p_{noise} = 0$, the estimator $\widehat{V}_{I,\text{tay}}$ exhibited a value of RB of about -4% for both simple random sampling without replacement

and Bernoulli sampling. The jackknife variance estimator performed well with values of RB equal to -2.1% for simple random sampling without replacement and 1.8% for Bernoulli sampling. However, for larger values of $\mathcal{K}$ both variance estimators did not perform well. For instance, for $\mathcal{K} = 33/150 \approx 0.29$, the estimator $\widehat{V}_{I,\text{tay}}$ underestimated the true variance with values of RB equal to -16.8% for simple random sampling without replacement and -14.6% for Bernoulli sampling. On the other hand, the estimator $\widehat{V}_{I,\text{jack}}$ was 20.9% too large for simple random sampling without replacement and 24.4% too large for Bernoulli sampling. Again, the magnitude of the bias increased significantly as $\mathcal{K}$ increased.

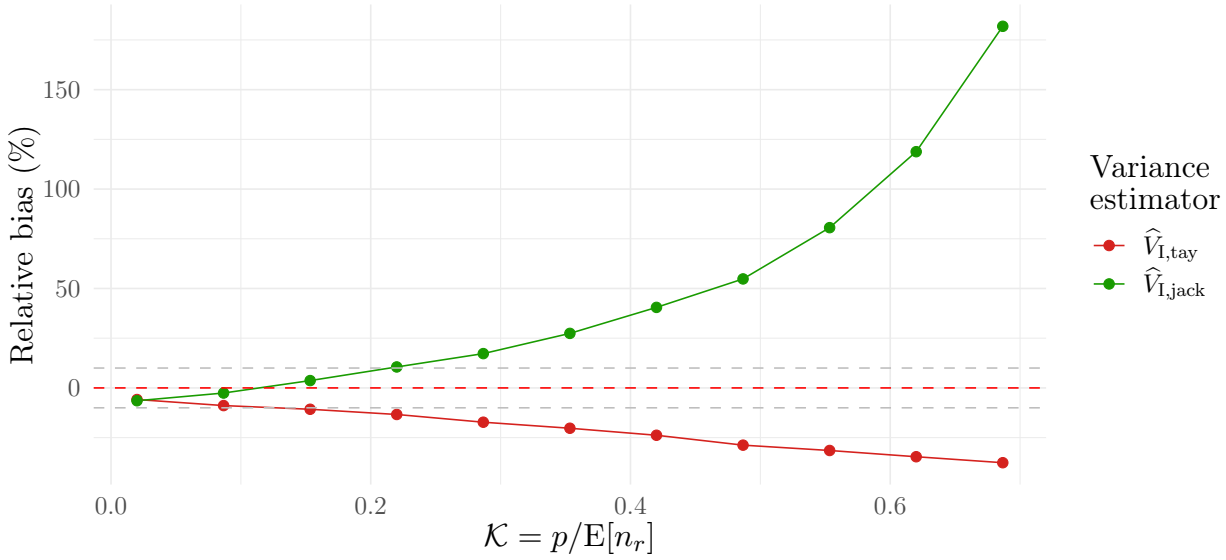

Figure 3: Behaviour of two variance estimators for $\widehat{\mu}_{lr}$ under simple random sampling without replacement.

3.3 Explaining the behavior of classical variance estimators

In the context of both model-assisted estimation and deterministic linear regression imputation, the customary variance estimators based on a first-order Taylor expansion and the generalized jackknife variance estimator tend to breakdown when $p/n \to \mathcal{K} \in (0, 1)$ (or $p/\mathbb{E}[n_r] \to \mathcal{K} \in (0, 1)$). In this section, we explain why this is the case. For simplicity, we confine to the case of model-assisted estimation under simple random sampling without

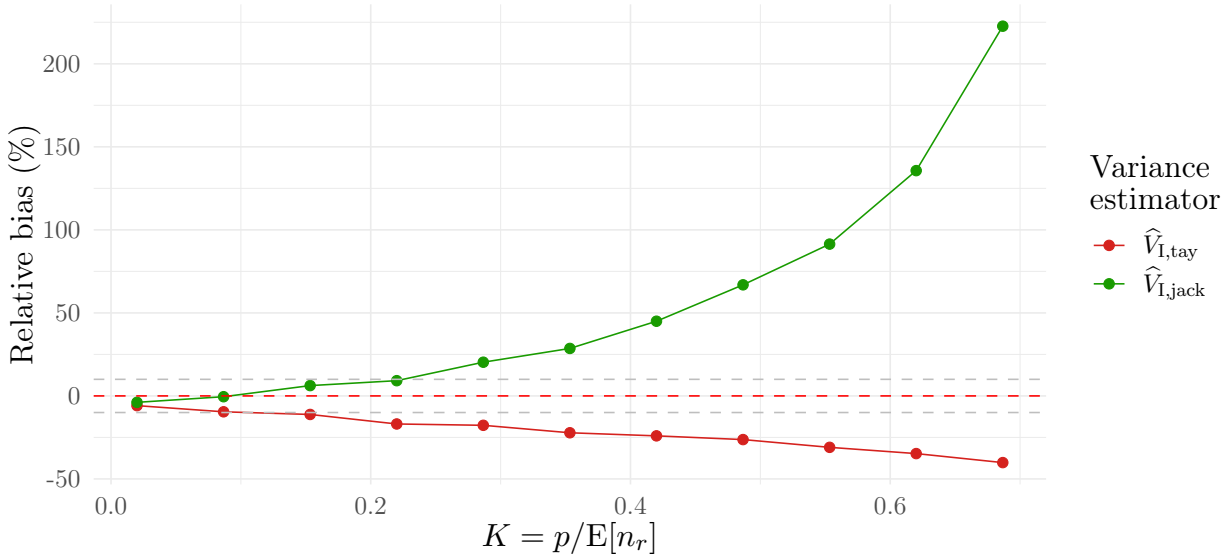

Figure 4: Behaviour of two variance estimators for $\widehat{\mu}_{lr}$ under Bernoulli sampling.

replacement. Arguments similar to the ones below can also be used to explain the behavior of classical variance estimators under deterministic linear regression imputation.

The variance estimator based on a first-order Taylor expansion given by (5) involves the sample residuals $\widehat{\epsilon}_{kS}$. It turns out that, in a high-dimensional setting, the distribution of the sample residuals $\widehat{\epsilon}_{kS}$ is not a good approximation of the distribution of the errors ϵ_k in (1). In particular, we have

$$\mathbb{V}_m(\widehat{\epsilon}_{kS}) = \sigma^2(1 - \widetilde{h}_{kk}), \qquad k \in S,$$

where $\widetilde{h}_{kk}$ denotes the kth diagonal element of the hat matrix $\boldsymbol{X}_S \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S^\top$. The validity of classical variance estimators relies on the assumption that $\widetilde{h}_{kk} \to 0$ as n and N go to infinity. In a high-dimensional setting, this assumption no longer holds, as it can be shown that

$$\widetilde{h}_{kk} = \frac{p}{n} + o_p(1), \qquad k \in S,$$

for a wide class of distributions for the design matrix $\boldsymbol{X}_S$ (e.g., the multivariate normal distribution); see e.g., El Karoui and Purdom (2018), Pajor and Pastur (2009) and Karoui and Koesters (2011) for a discussion. As a result, the variance of the sample residuals $\widehat{\epsilon}_{kS}$ is

approximately equal to $\sigma^2(1-p/n) \equiv \sigma^2(1-\mathcal{K})$, which can be considerably smaller than σ^2 for large values of $\mathcal{K}$. This, in turn, explains why the variance estimator based on a first-order Taylor expansion tends to underestimate the true variance of $\widehat{\mu}_{greg}$ for large values of $\mathcal{K}$.

Turning to generalized jackknife variance estimators, we note from (8) that it involves the residuals $\widehat{\epsilon}_{kS}^{(k)} = \widehat{\epsilon}_{kS}/(1-\widetilde{h}_{kk})$. Since

$$\mathbb{V}_m(\widehat{\epsilon}_{kS}^{(k)}) = \frac{\sigma^2}{1-\widetilde{h}_{kk}} \simeq \frac{\sigma^2}{1-\frac{p}{n}} \equiv \frac{\sigma^2}{1-\mathcal{K}}, \qquad k \in S,$$

the variance of $\widehat{\epsilon}_{kS}^{(k)}$ may be considerably larger than σ^2 for large values of $\mathcal{K}$. As a result, the generalized jackknife variance estimator tends to overestimate the true variance of $\widehat{\mu}_{greg}$ for large values of $\mathcal{K}$.

4 Bias: Model-assisted estimation

In this section, we provide a theoretical analysis on the bias of variance estimators for the GREG estimator in a high-dimensional setting. For simplicity, we confine to the case of Bernoulli sampling.

We consider the asymptotic framework of Isaki and Fuller (1982). We consider an increasing sequence of finite populations $\{U_v\}_{v\in\mathbb{N}}$ of sizes $\{N_v\}_{v\in\mathbb{N}}$ such that $U_v \subset U_{v+1}$, for all $v \in \mathbb{N}$. From U_v, a sample S_v is selected according to the sampling design $\mathcal{P}_v(\,\cdot\,, \mathbf{Z}_v)$. The first and second-order inclusion probabilities of $\mathcal{P}_v(\,\cdot\,, \mathbf{Z}_v)$ are denoted by $\{\pi_{k,v}\}_{k\in U_v}$ and $\{\pi_{k\ell,v}\}_{k\neq\ell\in U_v}$, respectively. For ease of notation, the subscript v will be omitted whenever possible. For two sequences $\{a_v\}_{v\in\mathbb{N}}$ and $\{b_v\}_{v\in\mathbb{N}}$, we write $a_v \simeq b_v$ to express that they share the same limit, i.e., $\lim_{v\to\infty} a_v/b_v = 1$. We extend this definition to sequences of random variables where the limit is to be understood in the probability sense. Moreover, asymptotic order notations are to be understood in a high-dimensional framework, whereby $\lim_{v\to\infty} p_v/n_v = \mathcal{K} \in (0;1)$.

In this article, the inference is made conditionally on the predictors. We assume the following regularity condition:

(H1) The design matrix $\boldsymbol{X}_S$ is such that, for all $k \in S$,

$$\widetilde{h}_{kk} = \frac{p_v}{n_v} + o(1).$$

Assumption (H1) should hold when the design matrix $\boldsymbol{X}_U$ corresponding to the population, is the realization of a random matrix; see Portnoy (1987) for more details, including a proof when $\boldsymbol{X}_U$ is a Gaussian data matrix.

Result 4.1. *Consider a Bernoulli sampling design and let* $\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)$ *be the variance of* $\widehat{\mu}_{greg,v}$ *with respect to the superpopulation model.*

i) If the superpopulation model is linear, the model variance $\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)$ *is unbiased for the unconditional variance* $\mathbb{V}\left(\widehat{\mu}_{greg,v}\right)$*, that is,*

$$\mathbb{E}_p\left[\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)\right] = \mathbb{V}\left(\widehat{\mu}_{greg,v}\right).$$

Moreover, if $\lim_{v\to\infty} \mathbb{V}_p\left(g_{k,v}\right) = 0$*, for all* k*, then* $\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)$ *and* $\mathbb{V}\left(\widehat{\mu}_{greg,v}\right)$ *are asymptotically equivalent, that is,*

$$\frac{\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)}{\mathbb{V}\left(\widehat{\mu}_{greg,v}\right)} \xrightarrow{\mathbb{P}} 1.$$

ii) The relative bias of $\widehat{V}_{\text{tay,v}}$*,* $\widehat{V}_{g,v}$ *and* $\widehat{V}_{jack,v}$*, given respectively by (5), (6) and (8), are given by*

$$\frac{\mathbb{E}_m\left[\widehat{V}_{\text{tay,v}}\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)} - 1 \simeq \frac{N}{\sum_{k\in U} g_k}\frac{n_s}{n}\left(1 + \pi\left\{\frac{n}{n_s} - 1\right\} - \mathcal{K}\left(1-\pi\right)\right) - 1, \tag{19}$$

$$\frac{\mathbb{E}_m\left[\widehat{V}_{g,v}\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)} - 1 \simeq \left(1-\pi\right)\left(1-\mathcal{K}\right) + \pi \times \frac{N}{\sum_{k\in U} g_k} - 1, \tag{20}$$

and

$$\frac{\mathbb{E}_m\left[\widehat{V}_{jack,v}\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)} - 1 \simeq \frac{1-\pi}{1-\mathcal{K}} + \pi \times \frac{N}{\sum_{k\in U} g_k} - 1. \tag{21}$$

Part i) of Result 4.1 shows that the model variance of the GREG estimator is unbiased and consistent for the unconditional variance provided that the linear regression model holds. Part i) also holds for general sampling designs, under appropriate assumptions on higher-order inclusion probabilities. Part ii) of Result 4.1 confirm the results of Section 3; that is, the variance estimators based on a first-order Taylor expansion lead to substantial under-estimation for large values of $\mathcal{K}$, whereas the variance estimator based on the generalized jackknife leads to substantial over-estimation of the true variance.

Remark 4.1. *The behavior of the g-weights* $\{g_k\}_{k\in U}$ *significantly impacts the high-dimensional behavior of the variance estimators. Indeed, if* p_v *is either fixed (or slowly increases with respect to* n_v*), it can be shown that, uniformly in* k*,*

$$g_k \xrightarrow{\mathbb{P}} 1.$$

This result may not hold for general covariates settings within a high-dimensional framework for which $\lim_{v\to\infty} p_v/n_v > 0$*. Note that the assumption in Result 4.1 that* $\lim_{v\to\infty} \mathbb{V}_p\left(g_{k,v}\right) = 0$ *is much weaker as it only requires that this limit is degenerate.*

Corollary 4.1. *Consider a Bernoulli sampling design with a negligible sampling fraction. Then, Expressions (19)-(21) reduce to*

$$\frac{\mathbb{E}_m\left[\widehat{V}_{\text{tay,v}}\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)} - 1 \simeq \frac{N}{\sum_{k\in U} g_k}(1-\mathcal{K}) - 1, \tag{22}$$

$$\frac{\mathbb{E}_m\left[\widehat{V}_{g,v}\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)} - 1 \simeq -\mathcal{K}, \tag{23}$$

and

$$\frac{\mathbb{E}_m\left[\widehat{V}_{jack,v}\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg,v}\right)} - 1 \simeq \frac{\mathcal{K}}{1-\mathcal{K}}, \tag{24}$$

respectively.

Interestingly, the bias of $\widehat{V}_{\text{jack}}$ matches the bias found in El Karoui and Purdom (2018) in the context of jackknife variance estimation of a prediction for a linear regression model.

We end this section by suggesting simple bias-adjusted variance estimators in the context of a high-dimensional setting. For a small sampling fraction, Expressions (22)-(24) motivates the following bias-adjusted version of $\widehat{V}_{\text{tay}}$, $\widehat{V}_g$ and $\widehat{V}_{\text{jack}}$. They are respectively given by

$$\widehat{V}_{\text{tay,v}}^{(adj)} = \frac{\sum_{k\in U} g_k}{N}(1-\mathcal{K})^{-1}\widehat{V}_{\text{tay,v}}, \tag{25}$$

$$\widehat{V}_{g,v}^{(adj)} = (1-\mathcal{K})^{-1}\widehat{V}_{g,v}, \tag{26}$$

and

$$\widehat{V}_{\text{jack,v}}^{(adj)} = (1-\mathcal{K})\widehat{V}_{\text{jack,v}}. \tag{27}$$

The bias-adjusted estimators (22)-(24) are asymptotically unbiased for all values of $\mathcal{K}$, and can be implemented using the data at hand.

5 Bias: Deterministic linear regression imputation

In this section, we study the behavior of several variance estimators in the context of deterministic linear regression imputation in a high-dimensional setting. To that aim, consider the following class of variance estimators:

$$\mathcal{V} := \Bigg\{ \widehat{V}^{(\psi)} := \frac{1}{\widehat{N}^2}\sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}\frac{\widehat{\mu}_{lr}-\widehat{\xi}_k^{(\psi)}}{\pi_k}\frac{\widehat{\mu}_{lr}-\widehat{\xi}_\ell^{(\psi)}}{\pi_\ell} + \frac{\sigma^2}{\widehat{N}^2}\sum_{k\in S_r}\frac{1}{\pi_k}\Big\{1-R_k(1+\widehat{\Gamma}_k)\Big\}^2;$$

$$\text{with} \quad \widehat{\xi}_\ell^{(\psi)} := \widetilde{y}_k + r_k\psi(\boldsymbol{X}_R)\widehat{\Gamma}_k\,\widehat{\epsilon}_{kR}, \quad \text{for some} \quad \psi : \mathbb{R}^{n_r\times p} \to \mathbb{R}\Bigg\}. \tag{28}$$

The class of variance estimators $\mathcal{V}$ includes, as special cases, the variance estimators (13) and (15) with $\psi(\boldsymbol{X}_R) = 1$ and $\psi(\boldsymbol{X}_R) = \widehat{h}_{kk}$, respectively. It also includes the variance estimator obtained by adjusting the residuals $\widehat{\epsilon}_{kR}$ using the rescaling factor $(1-\widehat{h}_{kk})^{1/2}$; see

e.g., El Karoui and Purdom (2018). This variance estimator, which we call the corrected variance estimator in the sequel, corresponds to the choice $\psi(\boldsymbol{X}_R) = (1 - \widehat{h}_{kk})^{-1/2}$. The rationale behind this choice is to recover the variance of the model errors ϵ_k.

The following result exhibits the asymptotic bias of any variance estimator belonging to the class $\mathcal{V}$.

Result 5.1. *Consider a Bernoulli sampling design and let $\widehat{V}^{(\psi)}$ be an arbitrary variance estimator belonging to the class $\mathcal{V}$ with functions $\{\psi_v\}_{v\in\mathbb{N}}$. Then,*

$$\frac{\mathbb{E}_m\left[\widehat{V}_v^{(\psi)}\right]}{\mathbb{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]} \backsimeq \frac{[1-\pi]\sum_{k\in S} B_k + \sigma^2 A_\psi}{[1-\pi]\sum_{k\in S} B_k + \sigma^2 A_{theo}},$$

where

$$A_\psi := (1-\pi)\left(n_r + \sum_{k\in S_m}\widehat{h}_{kk} + (1-\kappa)\left(\sum_{k\in S_r}\psi(\boldsymbol{X}_R)\Gamma_k\left\{2+\psi(\boldsymbol{X}_R)\Gamma_k\right\}\right)\right) + \pi\left(n_m + \sum_{k\in S_r}\Gamma_k^2\right),$$

$$A_{theo} := (1-\pi)(n_s - 1) + n_m + \sum_{k\in S_r}\Gamma_k^2,$$

with

$$B_k := \frac{1}{n_s}\sum_{\ell\in S}\left\{(\mathbf{x}_\ell^\top\boldsymbol{\beta})^2 - \mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta}\right\}. \tag{29}$$

In the case of a negligible sampling fraction, Corollary 5.1 below exhibits the expressions of the bias for the variance estimators (13) and (15) as well as the corrected variance estimator obtained with $\psi(\boldsymbol{X}_R) = (1 - \widehat{h}_{kk})^{-1/2}$ and denoted by $\widehat{V}_{\text{cor}}$.

Corollary 5.1. *Consider a Bernoulli sampling design with a negligible sampling fraction. Then,*

$$\frac{\mathbb{E}_m\left[\widehat{V}_{I,\text{tay,v}}\right]}{\mathbb{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]} - 1 \backsimeq \frac{\sum_{k\in S_m}\widehat{h}_{kk} - \mathcal{K}\left(2n_m + \sum_{k\in S_r}\Gamma_k^2\right)}{\sum_{k\in S}B_k/\sigma^2 + \left(n_r + 2n_m + \sum_{k\in S_r}\Gamma_k^2\right)}, \tag{30}$$

$$\frac{\mathbb{E}_m\left[\widehat{V}_{I,\text{jack,v}}\right]}{\mathbb{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]} - 1 \backsimeq \frac{\sum_{k\in S_m}\widehat{h}_{kk} + \frac{\mathcal{K}}{1-\mathcal{K}}\sum_{k\in S_r}\Gamma_k^2}{\sum_{k\in S}B_k/\sigma^2 + \left(n_r + 2n_m + \sum_{k\in S_r}\Gamma_k^2\right)} \tag{31}$$

and

$$\frac{\mathbb{E}_m\left[\widehat{V}_{I,\text{cor,v}}\right]}{\mathbb{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]} - 1 \backsimeq \frac{\sum_{k\in S_m}\widehat{h}_{kk} - 2n_m\left(1-\sqrt{1-\mathcal{K}}\right)}{\sum_{k\in S}B_k/\sigma^2 + \left(n_r + 2n_m + \sum_{k\in S_r}\Gamma_k^2\right)}. \tag{32}$$

Note that the terms on the right hand-side of (30)-(32) have the same denominator. Since $\sum_{k\in S}B_k \geq 0$, it follows that the sign of the bias in (30)-(32) depends on the sign of the numerator. In the case of $\widehat{V}_{I,\text{tay}}$, the sign of the numerator is expected to be negative, and the bias may be substantial for large values of $\mathcal{K}$. It follows from (31) that the bias of $\widehat{V}_{I,\text{jack}}$ is positive. Noting that $\mathcal{K}/(1-\mathcal{K})$ is monotonically increasing in $\mathcal{K} \in (0,1)$, the bias of $\widehat{V}_{I,\text{jack}}$ is expected to be large for large values of $\mathcal{K}$. Finally, looking at Expression (32), we expect the bias of $\widehat{V}_{I,\text{cor}}$ to be small to moderate as the term $1-\sqrt{1-\mathcal{K}}$ lies between 0 and 1.

Expressions (30)-(32) may be used to derive bias-adjusted versions of $\widehat{V}_{I,\text{tay}}$, and $\widehat{V}_{I,\text{jack}}$. Note that all the terms but B_k and σ^2 on the right hand-side of (30)-(32) can be computed using the data at hand. An estimator of B_k is obtained by replacing β in (29) by its weighted least square estimator given by (10). Also, a model-unbiased estimator of σ^2 is given by

$$\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{k\in S_r}(y_k - \mathbf{x}_k^\top\widehat{\boldsymbol{\beta}}_R)^2.$$

In Section 6.2, the bias-adjusted versions of $\widehat{V}_{I,\text{tay}}$ and $\widehat{V}_{I,\text{jack}}$ will be denoted by $\widehat{V}_{I,\text{jack}}^{(adj)}$ and $\widehat{V}_{I,\text{tay}}^{(adj)}$, respectively.

Remark 5.1. *The numerators on the right hand-side of (30)-(32) all include the term* $\sum_{k\in S_m}\widehat{h}_{kk}$, *which is a function of the predictors for both the respondents and the nonrespondents. Its magnitude heavily depends on the type of nonresponse mechanism, which is unknown. Indeed, when p is fixed, it can be shown that*

$$\sum_{k\in S_m}\widehat{h}_{kk} \simeq \lim_{v\to\infty}\frac{n_{m,v}}{n_{r,v}} \times tr\left(\mathbb{E}\left[\mathbf{x}_1\mathbf{x}_1^\top|R_1=1\right]^{-1}\mathbb{E}\left[\mathbf{x}_1\mathbf{x}_1^\top|R_1=0\right]\right).$$

As a result, unless the data are Missing Completely At Random, it is generally not possible to establish a general result about the magnitude of $\sum_{k\in S_m} \widehat{h}_{kk}$.

We end this section by determining the coefficients $\psi(\boldsymbol{X}_R)$ in (28) that produce an asymptotically unbiased variance estimator. This is given in the next result.

Result 5.2. *Consider a Bernoulli sampling design with a negligible sampling fraction. Let* $\widetilde{\mathcal{V}}$ *be the subset of* $\mathcal{V}$ *where the functions* ψ *are constants. The choices*

$$\psi_1(\boldsymbol{X}_R) := \frac{-B - \sqrt{B^2 - 4AC}}{2A} \tag{33}$$

and

$$\psi_2(\boldsymbol{X}_R) := \frac{-B + \sqrt{B^2 - 4AC}}{2A}, \tag{34}$$

with

$$A := (1-\kappa) \sum_{k\in S_r} \widehat{\Gamma}_k^2,$$

$$B := 2\,(1-\kappa)\, n_m,$$

$$C := \sum_{k\in S_m} \widehat{h}_{kk} - 2n_m - \sum_{k\in S_r} \widehat{\Gamma}_k^2,$$

produce two asymptotically unbiased variance estimators. Note that the roots $\psi_1(\boldsymbol{X}_R)$ and $\psi_1(\boldsymbol{X}_R)$ can be readily computed using the data at hand.

6 Empirical behavior of bias-adjusted estimators

In this section, we present the results from a simulation study that assessed the performance of several variance estimators and their associated bias-adjusted version in terms of relative bias. In Section 6.1, we consider the model-assisted estimation setup, while Section 6.2 discusses the linear regression imputation setup.

6.1 Bias-adjusted variance estimators: Model-assisted estimation

We used the same simulation setup as the one described in Section 3.1. To estimate the variance of $\widehat{\mu}_{greg}$, we computed several variance estimators in each sample: (i) $\widehat{V}_{\text{tay}}$ given by (5); (ii) $\widehat{V}_g$ given by (7); (iii) $\widehat{V}_{\text{jack}}$ given by (8); (iv) The bias-adjusted variance estimators $\widehat{V}_{\text{tay}}^{(adj)}$, $\widehat{V}_g^{(adj)}$ and $\widehat{V}_{\text{jack}}^{(adj)}$, given respectively by (25)-(27). In addition, we computed the variance $\mathbb{V}_m := \mathbb{V}_m(\widehat{\mu}_{greg})$, which corresponds to the model variance $\widehat{\mu}_{greg}$. For each variance estimator, we computed its Monte Carlo percent relative bias given by (18). The results for Bernoulli sampling are shown in Table 1, whereas Table 2 shows the results for simple random sampling without replacement.

From Table 1, we first note that observe that the target $\mathbb{V}_m$ exhibited a small bias for all values of $\mathcal{K}$, which is consistent with Result 4.1 part i). The results pertaining to $\widehat{V}_{\text{tay}}$, $\widehat{V}_g$ and $\widehat{V}_{\text{jack}}$ are virtually identical to those presented in Section 3.1. The bias-adjusted versions $\widehat{V}_{\text{tay}}^{(adj)}$, $\widehat{V}_g^{(adj)}$ and $\widehat{V}_{\text{jack}}^{(adj)}$ performed very well in terms of bias for all values of $\mathcal{K}$ with an absolute RB less than 7%. Very similar results (see Table 2) were obtained for simple random sampling without replacement, which can be explained by the fact that, in terms of variance, the strategy consisting of Bernoulli sampling and the GREG estimator is asymptotically equivalent to the strategy consisting of simple random sampling without replacement and the GREG estimator.

6.2 Bias-adjusted variance estimators: Linear regression imputation

We used the same simulation setup as the one described in Section 3.2. To estimate the variance of $\widehat{\mu}_{lr}$), we computed several variance estimators in each sample: (i) $\widehat{V}_{I,\text{tay}}$ given by (5); (ii) $\widehat{V}_{I,\text{jack}}$ given by (8); (iii) $\widehat{V}_{\text{cor}}$ obtained from (28) with $\psi(\boldsymbol{X}_R) = (1 - \widehat{h}_{kk})^{-1/2}$; (iv) The bias-adjusted variance estimators $\widehat{V}_{I,\text{tay}}^{(adj)}$, $\widehat{V}_{I,\text{jack}}^{(adj)}$; (v) The variance estimators obtained

			Relative bias					
p	$\mathcal{K}$	$\mathbb{V}_m$	$\widehat{V}_{\text{tay}}$	$\widehat{V}_g$	$\widehat{V}_{\text{jack}}$	$\widehat{V}_{\text{tay}}^{(adj)}$	$\widehat{V}_g^{(adj)}$	$\widehat{V}_{\text{jack}}^{(adj)}$
3	0.01	3.2	8.9	4.7	7.3	5.7	5.7	6.4
23	0.08	9.3	-9.0	-2.1	14.8	6.0	6.0	6.9
43	0.14	9.8	-20.8	-8.7	23.9	6.7	6.6	7.7
63	0.21	9.8	-31.8	-15.8	33.8	6.8	6.7	7.9
83	0.28	6.6	-44.0	-25.1	41.5	3.8	3.7	5.1
103	0.34	6.0	-53.5	-32.4	55.0	3.2	3.2	4.9
123	0.41	7.2	-61.8	-39.0	73.6	4.0	3.9	5.8
143	0.48	5.0	-70.1	-47.1	92.6	2.2	1.8	4.2
163	0.54	6.2	-76.6	-53.6	124.0	3.3	2.6	5.6
183	0.61	5.5	-82.8	-60.9	164.0	2.6	2.0	5.6
203	0.68	4.0	-88.1	-68.3	221.4	1.4	0.5	5.0

Table 1: Monte Carlo percent relative bias of several variance estimators for Bernoulli sampling: model-assisted estimation.

			Relative bias					
p	$\mathcal{K}$	$\mathbb{V}_m$	$\widehat{V}_{\text{tay}}$	$\widehat{V}_g$	$\widehat{V}_{\text{jack}}$	$\widehat{V}_{\text{tay}}^{(adj)}$	$\widehat{V}_g^{(adj)}$	$\widehat{V}_{\text{jack}}^{(adj)}$
3	0.01	5.4	4.2	5.2	7.9	6.2	6.2	6.9
23	0.08	6.4	-7.9	-1.1	16.1	7.2	7.0	8.1
43	0.14	4.1	-21.7	-10.2	21.7	4.9	4.8	5.9
63	0.21	3.5	-33.1	-17.6	30.8	4.3	4.2	5.6
83	0.28	3.5	-43.2	-24.4	42.5	4.6	4.4	6.1
103	0.34	4.3	-52.3	-30.9	57.6	5.3	5.1	7.1
123	0.41	2.2	-61.9	-39.1	71.5	3.1	3.0	5.3
143	0.48	4.6	-68.7	-44.8	97.6	5.7	5.4	8.1
163	0.54	3.5	-76.0	-52.3	124.3	4.5	4.3	7.5
183	0.61	1.5	-82.5	-60.2	157.4	2.3	1.9	5.8
203	0.68	-1.7	-88.0	-68.2	201.0	-1.2	-1.7	2.8

Table 2: Monte Carlo percent relative bias of several variance estimators for simple random sampling without replacement: model-assisted estimation.

from (28) with $\psi_1(\boldsymbol{X}_R)$ and $\psi_2(\boldsymbol{X}_R)$ given by (33) and (34), respectively, and denoted by $\widehat{V}_{\psi_1}^{(adj)}$ and $\widehat{V}_{\psi_2}^{(adj)}$. In addition, we computed the (unfeasible) target given by (41). Finally, we computed the (unfeasible) bias-adjusted variance estimators $\widetilde{V}_{I,\text{tay}}^{(adj)}$ and $\widetilde{V}_{I,\text{jack}}^{(adj)}$ that are identical to $\widehat{V}_{I,\text{tay}}^{(adj)}$ and $\widehat{V}_{I,\text{jack}}^{(adj)}$, respectively, except that they use the true value of B_k (see Equation (29)) and the true value of σ^2. For each variance estimator, we computed its Monte Carlo percent relative bias given by (18). The results for Bernoulli sampling are shown in

Table 3, whereas Table 4 shows the results for simple random sampling without replacement.

From Table 3, we first note that the target $\mathbb{V}_{target}$ exhibited a negligible bias for all values of $\mathcal{K}$, as expected. The results pertaining to $\widehat{V}_{I,\text{tay}}$ and $\widehat{V}_{I,\text{jack}}$ were virtually identical to those presented in Section 3.2. The variance estimator $\widehat{V}_{\text{cor}}$ performed much better than either $\widehat{V}_{I,\text{tay}}$ or $\widehat{V}_{I,\text{jack}}$, especially for small to moderate values of $\mathcal{K}$. For instance, its RB was equal to 2.9% for $\mathcal{K} = 0.42$. In contrast, the variance estimators $\widehat{V}_{I,\text{tay}}$ and $\widehat{V}_{I,\text{jack}}$ showed a RB equal to -22.7% and 45.1%, respectively for $\mathcal{K} = 0.42$. However, the performance of $\widehat{V}_{\text{cor}}$ slightly deteriorated for larger values of $\mathcal{K}$. For $\mathcal{K} = 0.69$, the value of RB was approximately equal to 17%. Turning to the bias-adjusted variance estimators $\widehat{V}_{I,\text{tay}}^{(adj)}$ and $\widehat{V}_{I,\text{jack}}^{(adj)}$, we note that they performed well for small to moderate values of $\mathcal{K}$. For instance, for $\mathcal{K} = 0.42$, the variance estimators $\widehat{V}_{I,\text{tay}}^{(adj)}$ and $\widehat{V}_{I,\text{jack}}^{(adj)}$ exhibited a value of RB of about -6.9% and 3.8%, respectively. However, for a large value of $\mathcal{K}$, we note a deterioration for both $\widehat{V}_{I,\text{tay}}^{(adj)}$ and $\widehat{V}_{I,\text{jack}}^{(adj)}$ with a value of RB of about -13.2% and 19.8%, respectively. This deterioration in terms of bias seems to correspond to the price to pay for estimating B_k in (29) in a high-dimensional setting. Indeed, if comparing $\widehat{V}_{I,\text{tay}}^{(adj)}$ and $\widehat{V}_{I,\text{jack}}^{(adj)}$ and their unfeasible versions $\widetilde{V}_{I,\text{tay}}^{(adj)}$ and $\widetilde{V}_{I,\text{jack}}^{(adj)}$ based of the true values of B_k, we note that the latter performed very well for all values of $\mathcal{K}$ with values of RB less than 4% in all the scenarios. More research is needed to derive estimators of B_k that would be more robust to the dimension of the vector of predictors $\mathbf{x}$. Finally, the variance estimators $\widehat{V}_{\psi_1}^{(adj)}$ and $\widehat{V}_{\psi_2}^{(adj)}$ performed very well in terms of bias for all values of $\mathcal{K}$. Both variance estimators were virtually unbiased in all the scenarios, which is consistent with Result 5.2. Again, the results for simple random sampling without replacement were similar to those obtained for Bernoulli sampling; see Table 4.

Relative bias										
$\mathcal{K}$	V_{target}	$\widehat{V}_{I,\text{tay}}$	$\widehat{V}_{I,\text{jack}}$	$\widehat{V}_{cor}$	$\widehat{V}^{(adj)}_{I,\text{tay}}$	$\widehat{V}^{(adj)}_{I,\text{jack}}$	$\widehat{V}^{(adj)}_{\psi_1}$	$\widehat{V}^{(adj)}_{\psi_2}$	$\widehat{V}^{(adj)}_{I,\text{tay}_\text{T}}$	$\widehat{V}^{(adj)}_{I,\text{jk}_\text{T}}$
0.02	1.2	-0.5	1.8	0.6	-1.6	-1.3	0.4	0.4	0.4	0.7
0.15	0.8	-7.7	10.4	0.7	-3.5	-1.9	-0.2	-0.1	-0.1	0.3
0.29	0.8	-14.6	24.4	2.0	-4.6	0.2	-0.4	-0.2	-0.2	0.4
0.42	0.8	-21.5	48.1	4.7	-6.9	3.8	-0.4	-0.3	-0.2	0.7
0.55	0.8	-28.6	91.0	8.9	-8.7	11.2	-0.6	-0.5	-0.4	1.0
0.69	0.9	-35.7	185.1	15.8	-13.2	19.8	-0.7	-0.6	-0.5	1.9

Table 3: Monte Carlo percent relative bias of several variance estimators for Bernoulli sampling: Linear regression imputation

Relative bias										
$\mathcal{K}$	V_{target}	$\widehat{V}_{I,\text{tay}}$	$\widehat{V}_{I,\text{jack}}$	$\widehat{V}_{cor}$	$\widehat{V}^{(adj)}_{I,\text{tay}}$	$\widehat{V}^{(adj)}_{I,jack}$	$\widehat{V}^{(adj)}_{\psi_1}$	$\widehat{V}^{(adj)}_{\psi_2}$	$\widehat{V}^{(adj)}_{I,\text{tay}_\text{T}}$	$\widehat{V}^{(adj)}_{I,\text{jk}_\text{T}}$
0.02	-2.8	-4.3	-2.1	-3.2	-0.3	-0.1	-3.5	-3.4	-3.4	-3.2
0.15	-3.1	-11.1	6.2	-3.1	-1.3	0.3	-4.0	-3.8	-3.8	-3.4
0.29	-2.0	-16.8	20.9	-0.7	-1.9	3.2	-3.1	-2.9	-2.8	-2.4
0.42	-1.0	-22.7	45.1	2.89	-5.0	5.8	-2.1	-1.9	-1.9	-0.9
0.55	1.1	-28.0	90.6	9.4	-8.0	12.0	-0.0	0.1	0.2	1.6
0.69	1.9	-34.6	182.1	17.0	-13.4	19.1	0.5	0.7	0.8	3.2

Table 4: Monte Carlo percent relative bias of several variance estimators for simple random sampling without replacement: Linear regression imputation

7 Final remarks

In this article, we studied the behavior of classical variance estimators in a high-dimensional setting. We considered two setups based on the customary linear regression model. Through a mix of theoretical and empirical investigations, we showed that customary variance estimators tend to break down as p/n increases. In the context of Bernoulli sampling and simple random sampling without replacement, we suggested bias-adjusted estimators that were shown to perform well, even in a high-dimensional setting. The extension of our results to the case of stratified simple random sampling without replacement is relatively straightforward, as the proposed variance estimators may be computed separately within each stratum.

Future work will involve extending our results in the case of generalized linear models, and high entropy sampling designs with inclusion probabilities proportional to size; e.g., see Berger

(1998, 2007) and Haziza et al. (2008) . Results not shown here suggest that the customary finite population bootstrap variance estimators (Mashreghi et al., 2016) suffer from substantial overestimation for large values of p/n. In the classical iid setup, El Karoui and Purdom (2018) and Zhao and Candes (2022) showed that classical bootstrap variance estimators, like the jackknife, suffer from substantial overestimation for large values of p/n. Developing bootstrap algorithms that lead to consistent variance estimators in the context of finite population sampling in a high-dimensional setting would be desirable. This will be treated elsewhere.

References

Berger, Y. G. (1998). Rate of convergence for asymptotic variance of the horvitz–thompson estimator. *Journal of Statistical Planning and Inference*, 74(1):149–168.

Berger, Y. G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94(4):953–964.

Berger, Y. G. and Rao, J. (2006). Adjusted jackknife for imputation under unequal probability sampling without replacement. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):531–547.

Berger, Y. G. and Skinner, C. J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):79–89.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205.

Campbell, C. (1980). A different view of finite population estimation. In Washington, D., editor, *Proc. Survey Res. Meth. Sect. Am. Statist. Assoc.*, pages 319–324.

Cardot, H., Goga, C., and Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27(243-260).

Chauvet, G. and Goga, C. (2022). Asymptotic efficiency of the calibration estimator in a high-dimensional data setting. *Journal of Statistical Planning and Inference*, 217:177–187.

Chen, S. and Haziza, D. (2019). Recent developments in dealing with item non-response in surveys: a critical review. *International Statistical Review*, 87:S192–S218.

Dagdoug, M., Goga, C., and Haziza, D. (2022). Model-assisted estimation in high-dimensional

settings for survey data. *Journal of Applied Statistics*. https://doi.org/10.1080/02664763.2022.2047905.

Duchesne, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16(2):133.

El Karoui, N. and Purdom, E. (2018). Can we trust the bootstrap in high-dimensions? the case of linear models. *The Journal of Machine Learning Research*, 19(1):170–235.

Fay, R. (1991). *A design-based perspective on missing data variance*. US Census Bureau.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C., editors, *Handbook of statistics*, volume 29A, pages 215–246. Elsevier.

Haziza, D., Mecatti, F., and Rao, J. (2008). Evaluation of some approximate variance estimators under the rao-sampford unequal probability sampling design. *Metron*, 66(1):91–108.

Haziza, D. and Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data: a critical review. *Japanese Journal of Statistics and Data Science*, 3(2):583–623.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, 77:49–61.

Karoui, N. E. and Koesters, H. (2011). Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *arXiv: Statistics Theory*.

Kim, J. K. and Rao, J. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96(4):917–932.

Mashreghi, Z., Haziza, D., and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10(none).

Pajor, A. and Pastur, L. (2009). On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution. *Studia Mathematica*, 1(195):11–29.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In *Handbook of statistics*, volume 29, pages 455–487. Elsevier.

Portnoy, S. (1987). A central limit theorem applicable to robust regression estimators. *Journal of multivariate analysis*, 22(1):24–50.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Särndal, C. E. (1980). On pi-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67(3):639–650.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18:241–252.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33:99–119.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3):527–537.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94(445):254–265.

Ta, T., Shao, J., Li, Q., and Wang, L. (2020). Generalized regression estimators with high-dimensional covariates. *Statistica Sinica*.

Valliant, R. (2002). Variance estimation for the general regression estimator. *Survey methodology*, 28(1):103–108.

Wolter, K. (2007). *Introduction to Variance Estimation*. New York: Springer.

Yung, W. and Rao, J. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22:23–32.

Zhao, Q. and Candes, E. J. (2022). An adaptively resized parametric bootstrap for inference in high-dimensional generalized linear models. *arXiv preprint arXiv:2208.08944*.

Appendices

A Proof of Proposition 2.1.

The generalized jackknife variance estimator of $\widehat{\mu}_{greg}$ is defined in Berger and Skinner (2005) as

$$\widehat{V}_{\text{jack}}\left(\widehat{\mu}_{greg}\right) := \sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}z_k z_\ell, \tag{35}$$

where

$$z_k := (1-w_k)\left(\widehat{\mu}_{greg} - \widehat{\mu}_{greg}^{(k)}\right). \tag{36}$$

In (36), $\widehat{\mu}_{greg}^{(k)}$ denotes the GREG estimator as defined in (2), computed by deleting the element k of the sample (but not of the population). More generally, in what follows, we use the subscript (k) to denote any statistic (or, set) computed after deleting element k.

To derive a closed-form formula for (35), it suffices to find a closed-form formula for $\widehat{\mu}_{greg}^{(k)}$. In Duchesne (2000), a formula is given when the inclusion probabilities are reweighted after the deletion of an element, a practice that we do not do with the generalized jackknife variance estimator. Below, we give a simple proof relying on the following lemma.

Lemma 1. *The following relation holds:*

$$\widehat{\boldsymbol{\beta}}^{(k)} = \widehat{\boldsymbol{\beta}} - \frac{\boldsymbol{A}_{\Pi S}^{-1}d_k\mathbf{x}_k\widehat{\epsilon}_{kS}}{1-\widetilde{h}_{kk}^{\pi}}. \tag{37}$$

Lemma 1 is a weighted extension of the well-known closed-form formula for leave-one-out linear regression. We let $\widehat{t}_{greg} := N\widehat{\mu}_{greg}$ be the usual total GREG estimator. We begin by noting that if the full-sample GREG estimator can be written in projection form, then the leave-one-out version of the estimator can, too. It follows that

$$\begin{aligned}\widehat{t}_{greg}^{(k)} &= \boldsymbol{t}_x^\top\widehat{\boldsymbol{\beta}}^{(k)}\\ &= \boldsymbol{t}_x^\top\left(\widehat{\boldsymbol{\beta}} - \frac{\boldsymbol{A}_{\Pi S}^{-1}d_k\mathbf{x}_k\widehat{\epsilon}_{kS}}{1-\widetilde{h}_{kk}^{\pi}}\right)\end{aligned}$$

$$= \widehat{t}_{greg} - \frac{d_k g_k \widehat{\epsilon}_{ks}}{1 - \widetilde{h}_{kk}^{\pi}}.$$

Now,

$$\widehat{\mu}_{greg}^{(k)} := \frac{\widehat{t}_{greg}^{(k)}}{N} = \widehat{\mu}_{greg} - \frac{d_k g_k \widehat{\epsilon}_{ks}}{N\left(1 - \widetilde{h}_{kk}^{\pi}\right)},$$

from which it follows that

$$\widehat{\mu}_{greg}^{(k)} - \widehat{\mu}_{greg} = -\frac{1}{N}\frac{d_k g_k \widehat{\epsilon}_{ks}}{1 - \widetilde{h}_{kk}^{\pi}}.$$

Finally,

$$z_k = \frac{1}{N}\frac{(1 - w_k)\, g_k \widehat{\epsilon}_{ks}}{\pi_k\left(1 - \widetilde{h}_{kk}^{\pi}\right)},$$

which concludes the proof.

B Proof of Result 2.1.

The Generalized Jackknife variance estimator proposed in Berger and Rao (2006) is

$$\widehat{V}_{\text{I,jack}}\left(\widehat{\mu}_{lr}\right) := \sum_{k \in S}\sum_{\ell \in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}} e_k e_\ell, \tag{38}$$

where

$$e_k := (1 - w_k)\left(\widehat{\mu}_{lr} - \widehat{\mu}_{lr}^{(k)}\right).$$

The effect of deleting a respondent or nonrespondent is different; we treat these cases separately. For $k \in S_m$, write

$$\widehat{\mu}_{lr}^{(k)} := \frac{1}{\widehat{N} - d_k}\left(\sum_{\ell \in S_r}\frac{y_\ell}{\pi_\ell} + \sum_{j \in S_m^{(k)}}\frac{\mathbf{x}_\ell^\top \widehat{\boldsymbol{\beta}}_R}{\pi_\ell}\right)$$

$$= \frac{1}{\widehat{N} - d_k}\left(\sum_{\ell \in S_r}\frac{y_\ell}{\pi_\ell} + \sum_{\ell \in S_m}\frac{\mathbf{x}_\ell^\top \widehat{\boldsymbol{\beta}}_R}{\pi_\ell} - \frac{\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_R}{\pi_k}\right)$$

$$= \frac{\widehat{N}}{\widehat{N} - d_k}\left(\widehat{\mu}_{lr} - \frac{\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_R}{\widehat{N}\pi_k}\right).$$

Similarly, for $k \in S_r$, we have

$$\widehat{\mu}_{lr}^{(k)} := \frac{1}{\widehat{N} - d_k} \left(\sum_{\ell \in S_r^{(k)}} \frac{y_\ell}{\pi_\ell} + \sum_{\ell \in S_m} \frac{\mathbf{x}_\ell^\top \widehat{\boldsymbol{\beta}}_R^{(k)}}{\pi_\ell} \right)$$

$$= \frac{1}{\widehat{N} - d_k} \left\{ \sum_{\ell \in S_r} \frac{y_\ell}{\pi_\ell} - \frac{y_k}{\pi_k} + \sum_{\ell \in S_m} \frac{\mathbf{x}_\ell^\top}{\pi_\ell} \left(\widehat{\boldsymbol{\beta}}_r - \frac{\mathbf{A}_{\Pi R}^{-1} \mathbf{x}_k \widehat{\epsilon}_{Rk}}{\pi_k (1 - \widehat{h}_{kk}^{\pi})} \right) \right\}$$

$$= \frac{\widehat{N}}{\widehat{N} - d_k} \left[\widehat{\mu}_{lr} - \frac{1}{\widehat{N} \pi_k} \left\{ y_k + \sum_{\ell \in S_m} \frac{\mathbf{x}_\ell^\top}{\pi_\ell} \left(\frac{d_k \mathbf{A}_{\Pi R}^{-1} \mathbf{x}_k \widehat{\epsilon}_{Rk}}{1 - \widehat{h}_{kk}^{\pi}} \right) \right\} \right]$$

Thus, introducing response indicators, we obtain for an arbitrary element $k \in S$,

$$\widehat{\mu}_{lr}^{(k)} = \frac{\widehat{N}}{\widehat{N} - d_k} \left(\widehat{\mu}_{lr} - \frac{1}{\widehat{N} \pi_k} \widehat{\xi}_k^{(jk)} \right)$$

from which it follows that

$$\widehat{\mu}_{lr}^{(k)} - \widehat{\mu}_{lr} = \frac{d_k}{\widehat{N} - d_k} \left(\widehat{\mu}_{lr} - \widehat{\xi}_k^{(jk)} \right).$$

Therefore, a, closed-form formula of e_k is given by

$$e_k = \frac{d_k}{\widehat{N}} \left(\widehat{\mu}_{lr} - \widehat{\xi}_k^{(jk)} \right).$$

Replacing e_k by its closed formula in (38) leads to the result.

C Proof of Result 4.1

Statement i)

Consider the following variance decomposition:

$$\mathbb{V}\left(\widehat{\mu}_{greg}\right) = \mathbb{E}_p\left[\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right] + \mathbb{V}_p\left(\mathbb{E}_m\left[\widehat{\mu}_{greg}\right]\right).$$

Note that,

$$\mathbb{E}_m\left[\widehat{\mu}_{greg}\right] = \mathbb{E}_m\left[\frac{1}{N} \sum_{k \in U} \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_S\right] = \frac{1}{N} \sum_{k \in U} \mathbf{x}_k^\top \boldsymbol{\beta}.$$

In particular, observe that this quantity is independent of S, so that

$$\mathbb{V}\left(\widehat{\mu}_{greg}\right) = \mathbb{E}_p\left[\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right],$$

when the model is correctly specified. This establishes unbiasedness. For the asymptotic equivalence, we show that

$$\lim_{v\to\infty} n_v^2 \times \mathbb{E}\left[\left\{\mathbb{V}\left(\widehat{\mu}_{greg}\right) - \mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right\}^2\right] = 0.$$

Write $\widehat{\mu}_{greg}$ as $\widehat{\mu}_{greg} = \boldsymbol{\mu}_x^\top \widehat{\boldsymbol{\beta}}_S$ with $\boldsymbol{\mu}_x := \boldsymbol{t}_x/N$. It follows that

$$\mathbb{V}_m\left(\widehat{\mu}_{greg}\right) = \mathbb{V}_m\left(\boldsymbol{\mu}_x^\top \widehat{\boldsymbol{\beta}}_S\right) = \boldsymbol{\mu}_x^\top \mathbb{V}_m\left(\widehat{\boldsymbol{\beta}}_S\right)\boldsymbol{\mu}_x = \frac{\sigma^2}{\pi}\boldsymbol{\mu}_x^\top \boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{\mu}_x = \frac{\sigma^2}{\pi N^2}\sum_{k\in U} g_k.$$

We begin with the following decomposition:

$$\mathbb{V}\left(\widehat{\mu}_{greg}\right) - \mathbb{V}_m\left(\widehat{\mu}_{greg}\right) = \frac{\sigma^2}{\pi N_v^2}\sum_{k\in U_v}\left(g_{k,v} - \mathbb{E}_p\left[g_{k,v}\right]\right).$$

It thus follows that

$$\begin{aligned} n_v^2 \times \mathbb{E}_p\left[\left\{\mathbb{V}\left(\widehat{\mu}_{greg}\right) - \mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right\}^2\right] &= \frac{n_v^2\sigma^4}{\pi^2 N_v^4}\mathbb{E}_p\left[\left\{\sum_{k\in U_v}\left(g_{k,v} - \mathbb{E}_p\left[g_{k,v}\right]\right)\right\}^2\right] \\ &\leqslant \frac{n_v^2\sigma^4}{\pi^2 N_v^3}\sum_{k\in U_v}\mathbb{E}_p\left[\left\{\left(g_{k,v} - \mathbb{E}_p\left[g_{k,v}\right]\right)\right\}^2\right]. \end{aligned}$$

By symmetry, we obtain

$$n_v^2 \times \mathbb{E}_p\left[\left\{\mathbb{V}\left(\widehat{\mu}_{greg}\right) - \mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right\}^2\right] \leqslant \frac{n_v^2\sigma^4}{\pi^2 N_v^2}\mathbb{V}_p\left(g_{1,v}\right) = o(1),$$

by assumption.

Statement ii)

Taylor: Under a Bernoulli sampling design, the Taylor variance estimator reduces to

$$\widehat{V}_{\text{tay}} = \frac{1}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\widehat{\epsilon}_{kS}^2 + \frac{\widehat{\sigma}^2}{N}.$$

Thus,

$$\mathbb{E}_m\left[\widehat{V}_{\text{tay}}\right] = \frac{1}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\sigma^2\left(1-\widetilde{h}_{kk}\right)+\frac{\sigma^2}{N}$$

$$= \frac{\sigma^2}{N^2}\frac{1-\pi}{\pi^2}\left(n_s-p\right)+\frac{\sigma^2}{N}$$

$$= \frac{\sigma^2 n_s}{N^2}\frac{1-\pi}{\pi^2}\left(1-\mathcal{K}\right)+\frac{\sigma^2}{N}.$$

It follows that

$$\frac{\mathbb{E}_m\left[\widehat{V}_{\text{tay}}\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} = \frac{n_s\left(1-\pi\right)\pi^{-1}\left(1-\mathcal{K}\right)}{\sum_{k\in U}g_k}+\frac{\pi N}{\sum_{k\in U}g_k}$$

$$= \frac{n_s}{n}\times\left(1-\pi\right)\left(1-\mathcal{K}\right)\times\frac{N}{\sum_{k\in U}g_k}+\frac{\pi N}{\sum_{k\in U}g_k}$$

$$= \frac{N}{\sum_{k\in U}g_k}\frac{n_s}{n}\left(\left(1-\pi\right)\left(1-\mathcal{K}\right)+\frac{n}{n_s}\pi\right)$$

$$= \frac{N}{\sum_{k\in U}g_k}\frac{n_s}{n}\left(1+\pi_v\left\{\frac{n}{n_s}-1\right\}-\mathcal{K}\left(1-\pi\right)\right).$$

G-weighted: We begin by noting the following fact:

$$\sum_{k\in S}g_k^2 = N^2\pi^2\sigma^{-2}\mathbb{V}_m\left(\widehat{\mu}_{greg}\right). \tag{39}$$

To prove (39), write

$$\sum_{k\in S}g_k^2 = \sum_{k\in S}\boldsymbol{t}_{\mathbf{x}}\boldsymbol{A}_{\Pi S}^{-1}\mathbf{x}_k\mathbf{x}_k\boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{t}_{\mathbf{x}}$$

$$= \pi\times\boldsymbol{t}_{\mathbf{x}}\boldsymbol{A}_{\Pi S}^{-1}\sum_{k\in S}\frac{\mathbf{x}_k\mathbf{x}_k}{\pi}\boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{t}_{\mathbf{x}}$$

$$= \pi\times\boldsymbol{t}_{\mathbf{x}}\boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{t}_{\mathbf{x}}$$

$$= N^2\pi^2\sigma^{-2}\times\frac{\sigma^2}{N^2\pi}\boldsymbol{t}_{\mathbf{x}}\boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{t}_{\mathbf{x}}$$

$$= N^2\pi^2\sigma^{-2}\mathbb{V}_m\left(\widehat{\mu}_{greg}\right).$$

Moreover,

$$\mathbb{E}_m\left[\widehat{V}_{\text{g}}\right] = \frac{1}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\sigma^2\left(1-\widetilde{h}_{kk}\right)g_k^2+\frac{\sigma^2}{N}$$

$$\simeq \frac{1-\mathcal{K}}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\sigma^2 g_k^2 + \frac{\sigma^2}{N}$$

$$= (1-\mathcal{K})(1-\pi)\,\mathbb{V}_m\left(\widehat{\mu}_{greg}\right) + \frac{\sigma^2}{N}.$$

Thus,

$$\frac{\mathbb{E}_m\left[\widehat{V}_{\mathrm{g}}\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} \simeq \frac{(1-\mathcal{K})(1-\pi)\,\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} + \frac{\pi N}{\sum_{k\in U} g_k}$$

$$= (1-\mathcal{K})(1-\pi) + \frac{\pi N}{\sum_{k\in U} g_k}.$$

Jackknife: In Bernoulli sampling, the Jackknife variance estimator reduces to

$$\widehat{V}_{\mathrm{jack}} - \frac{\widehat{\sigma}^2}{N} = \frac{(n-1)^2}{n^2N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\frac{g_k^2\widehat{\epsilon}_{ks}^2}{\left(1-\widetilde{h}_{kk}\right)^2} \simeq \frac{1}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\frac{g_k^2\,\widehat{\epsilon}_{ks}^2}{\left(1-\widetilde{h}_{kk}\right)^2},$$

so that

$$\mathbb{E}_m\left[\widehat{V}_{\mathrm{jack}}\right] - \frac{\sigma^2}{N} = \frac{(n-1)^2}{n^2N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\frac{g_k^2\widehat{\epsilon}_{ks}^2}{\left(1-\widetilde{h}_{kk}\right)^2} \simeq \frac{\sigma^2}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\frac{g_k^2}{\left(1-\widetilde{h}_{kk}\right)}.$$

Using (39), we get

$$\mathbb{E}_m\left[\widehat{V}_{\mathrm{jack}}\right] \simeq \frac{1-\pi}{1-K}\mathbb{V}_m\left(\widehat{\mu}_{greg}\right) + \frac{\sigma^2}{N}.$$

It follows that

$$\frac{\mathbb{E}_m\left[\widehat{V}_{\mathrm{jack}}\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} \simeq \frac{1-\pi}{1-K} + \frac{\pi N}{\sum_{k\in U} g_k}.$$

D Proof of Result 5.1.

Step 1: Derivation of a target.

We first obtain an expression of the target variance of $\widehat{\mu}_{lr}$ using the method of Särndal (1992).

This estimator, denoted $V_{target}(\widehat{\mu}_{lr})$, will remain unbiased even in cases where $p_v/n_v \to \mathcal{K} > 0$.

Using the law of total variance on $\widehat{\mu}_{lr}$, we get the following decomposition:

$$\mathbb{V}(\widehat{\mu}_{lr}) = \mathbb{V}_{sam}(\widehat{\mu}_{lr}) + \mathbb{V}_{nr}(\widehat{\mu}_{lr}) + \mathbb{V}_{mix}(\widehat{\mu}_{lr}),$$

where

$$\mathbb{V}_{sam}(\widehat{\mu}_{lr}) := \mathbb{E}_m \mathbb{V}_p(\widehat{\mu}_H),$$

$$\mathbb{V}_{nr}(\widehat{\mu}_{lr}) = \mathbb{E}_q \mathbb{E}_p \mathbb{V}_m(\widehat{\mu}_{lr} - \widehat{\mu}_H),$$

and

$$\mathbb{V}_{\text{mix}}(\widehat{\mu}_{lr}) := 2\mathbb{E}_p \mathbb{E}_q \text{Cov}_m \left\{\widehat{\mu}_H - \mu_y, \widehat{\mu}_{lr} - \widehat{\mu}_H\right\},$$

and $\widehat{\mu}_H := \widehat{N}^{-1} \sum_{k \in S} \pi_k^{-1} y_k$. In the case of linear regression imputation and Bernoulli sampling, it can be shown that $\mathbb{V}_{mix}(\widehat{\mu}_{lr}) = 0$. Now, using a first-order Taylor expansion, a full sample estimator of $\mathbb{V}_{sam}(\widehat{\mu}_{lr})$ is given by

$$\widehat{V}_{sam}(\widehat{\mu}_{lr}) = \frac{1-\pi}{n_s^2} \sum_{k \in S} (y_k - \widehat{\mu}_H)^2 .$$

We now turn to the nonresponse component. It can be shown that

$$\mathbb{V}_{nr} = \mathbb{E}_q \mathbb{E}_p \left[\frac{\sigma^2}{(\widehat{N}\pi)^2} \sum_{k \in S} \left\{ R_k (1 + \widehat{\Gamma}_k) - 1 \right\}^2 \right]. \tag{40}$$

Assuming σ^2 is known, the above quantity can be estimated by

$$\widehat{V}_{nr} = \frac{\sigma^2}{n_s^2} \widehat{A}_n,$$

where

$$\widehat{A}_n = \sum_{k \in S_r} (1 + \widehat{\Gamma}_k)^2 - n_s = n_m + \sum_{k \in S_r} \widehat{\Gamma}_k^2.$$

The target variance estimator is therefore given by

$$\widehat{V}_{target} := \widehat{V}_{sam} + \widehat{V}_{nr} = \frac{1-\pi}{n_s^2} \sum_{k \in S} (y_k - \widehat{\mu}_H)^2 + \frac{\sigma^2}{n_s^2} \widehat{A}_n, \tag{41}$$

Expanding the square and taking model expectations give

$$\mathbb{E}_m \left[\widehat{V}_{sam}(\widehat{\mu}_{lr}) \right] = \frac{1-\pi}{n_s^2} \sum_{k \in S} \mathbb{E}_m \left[(y_k - \widehat{\mu}_H)^2 \right]$$

$$= \frac{1-\pi}{n_s^2}\left\{\sum_{k\in S}(\mathbf{x}_k^\top\boldsymbol{\beta})^2 + n_s\sigma^2 - \frac{2}{n_s}\sum_{k\in S}\sum_{\ell\in S}\mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta} - 2\sigma^2 + \frac{1}{n}\sum_{k\in S}\sum_{\ell\in S}\mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta} + \sigma^2\right\}$$

$$= \frac{1-\pi}{n_s^2}\left\{\sum_{k\in S}B_k + \sigma^2\left(n_s - 1\right)\right\}, \tag{42}$$

where

$$B_k = \frac{1}{n_s}\sum_{\ell\in S}\left\{(\mathbf{x}_\ell^\top\boldsymbol{\beta})^2 - \mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta}\right\}.$$

Adding (42) and (40) and ignoring the negligible terms, we obtain an unbiased target variance estimator of $\mathbb{E}_m[V_{target}(\widehat{\mu}_{lr})]$

$$\mathbb{E}_m\left[V_{target}(\widehat{\mu}_{lr})\right] = \mathbb{E}_m\left(\widehat{V}_{sam} + \widehat{V}_{nr}\right) = \frac{1}{n_s^2}\left\{(1-\pi)\sum_{k\in S}B_k + \sigma^2\left((1-\pi)\,n_s + n_m + \sum_{k\in S_r}\widehat{\Gamma}_k^2\right)\right\}.$$

Step 2: Computing the model expectation of an arbitrary estimator in $\mathcal{V}$.

Let $\widehat{V}^{(\psi)}$ be an arbitrary estimator in $\mathcal{V}$. Since $\widehat{V}^{(\psi)} \in \mathcal{V}$, there exists some $\{\psi_v\}_{v\in\mathbb{N}}$ satisfying

$$\widehat{V}^{(\psi)} = \frac{1}{\widehat{N}^2}\sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}\frac{\widehat{\mu}_{lr} - \widehat{\xi}_k^{(\psi_v)}}{\pi_k}\frac{\widehat{\mu}_{lr} - \widehat{\xi}_\ell^{(\psi_v)}}{\pi_\ell} + \frac{\sigma^2}{\widehat{N}^2}\sum_{k\in S_r}\frac{1}{\pi_k}\left\{1 - R_k(1+\widehat{\Gamma}_k)\right\}^2, \tag{43}$$

with

$$\widehat{\xi}_k^{(\psi_v)} := \widetilde{y}_k + r_k\psi_v(\boldsymbol{X}_R)\widehat{\Gamma}_k\,\widehat{\epsilon}_{kR}, \qquad k \in S.$$

In case of Bernoulli sampling, $\widehat{V}^{(\psi)}$ in (43) reduces to

$$\widehat{V}^{(\psi)} = \frac{1-\pi}{n_s^2}\sum_{k\in S}\left(\widehat{\mu}_{lr} - \widehat{\xi}_k^{(\psi_v)}\right)^2 + \frac{\pi\sigma^2}{n_s^2}\left(n_m + \sum_{k\in S_r}\widehat{\Gamma}_k^2\right).$$

Remark D.1. *Without loss of generality, we assume here that σ^2 for simplicity; we may easily replace it with $\widehat{\sigma}^2 := (n_r - p)^{-1}\sum_{k\in S_r}\widehat{\epsilon}_{kR}^2$, which is unbiased for σ^2, independently of the number of covariates.*

We start by expanding the square as follows:

$$\mathbb{E}_m\left[\left(\widehat{\mu}_{lr}-\widehat{\xi}_k^{(\psi_v)}\right)^2\right]=\mathbb{E}_m\left[\widehat{\mu}_{lr}^2\right]-2\mathbb{E}_m\left[\widehat{\mu}_{lr}\widehat{\xi}_k^{(\psi_v)}\right]+\mathbb{E}_m\left[\left(\widehat{\xi}_k^{(\psi_v)}\right)^2\right]:=A_1(k)-2A_2(k)+A_3.$$

After some tedious though relatively straightforward algebra, it can be shown that

$$A_1(k)=\left(\mathbf{x}_k^\top\boldsymbol{\beta}\right)^2+\sigma^2\left(R_k+[1-R_k]\,\widehat{h}_{kk}+2R_k\widehat{\Gamma}_k\psi_v(\boldsymbol{X}_R)\left[1-\widehat{h}_{kk}\right]+R_k\widehat{\Gamma}_k^2\psi_v(\boldsymbol{X}_R)^2\left[1-\widehat{h}_kk\right]\right)$$

$$A_2(k)=\frac{1}{n_s}\left(\mathbf{x}_k^\top\boldsymbol{\beta}\sum_{\ell\in S}\mathbf{x}_l^\top\boldsymbol{\beta}+\sigma^2\left(1+\widehat{\Gamma}_k\right)\right)$$

$$A_3=\frac{1}{n_s^2}\left(\left[\sum_{k\in S}\mathbf{x}_k^\top\boldsymbol{\beta}\right]^2+\sigma^2\left[n_r+2n_m+\sum_{k\in S_m}\widehat{\Gamma}_k\right]\right).$$

Summing over k and ignoring the negligible terms, we obtain the following asymptotic equivalence

$$\mathbb{E}_m\left[\sum_{k\in S}\left(\widehat{\mu}_{lr}-\widehat{\xi}_k^{(\psi_v)}\right)^2\right]\simeq\sum_{k\in S}B_k$$

$$+\sigma^2\Bigg\{(1-\pi)\left(n_r+\sum_{k\in S_m}\widehat{h}_{kk}+(1-\kappa)\left(\sum_{k\in S_r}\psi(\boldsymbol{X}_R)\Gamma_k\left\{2+\psi(\boldsymbol{X}_R)\Gamma_k\right\}\right)+\pi\left(n_m+\sum_{k\in S_r}\Gamma_k^2\right)\right).$$

Taking the ratios of $\mathbb{E}_m\left[\widehat{V}^{(\psi)}\right]$ and $\mathbb{E}_m\left[V_{target}(\widehat{\mu}_{lr})\right]$ yields the result.

E Proof of Result 5.2

We let $\widehat{V}^{(\psi)}$ be an arbitrary element of $\widetilde{\mathcal{V}}$ with constant $\psi\in\mathbb{R}$. Next, using Result 5.1 and Corollary 5.1, we write the absolute asymptotic relative bias of $\widehat{V}$ to get

$$\text{ARB}\left(\widehat{V}^{(\psi)}\right):=\frac{\left|\mathbb{E}_m\left[\widehat{V}^{(\psi)}\right]-\mathbb{E}_m\left[\widehat{V}_{target}\right]\right|}{\mathbb{E}_m\left[\widehat{V}_{target}\right]}=\frac{\sigma^2\left|A_\psi-A_{theo}\right|}{\mathbb{E}_m\left[\widehat{V}_{target}\right]},\tag{44}$$

where it can be shown that, in our setting,

$$A_\psi:=n_r+\sum_{k\in S_m}\widehat{h}_{kk}+(1-\kappa)\left(2\psi n_m+\psi^2\sum_{k\in S_r}\widehat{\Gamma}_k\right),$$

$$A_{theo} := n_r + 2n_m + \sum_{k \in S_r} \Gamma_k^2.$$

Clearly, the absolute asymptotic relative bias takes only positive values; in what follows, we view this quantity as a function of ψ over the real line. Hence,

$$\min_{\psi \in \mathbb{R}} \frac{\left| \mathbb{E}_m \left[\widehat{V}^{(\psi)} \right] - \mathbb{E}_m \left[\widehat{V}_{target} \right] \right|}{\mathbb{E}_m \left[\widehat{V}_{target} \right]} \geqslant 0.$$

Proving the existence of zeros of the numerator is therefore sufficient to find minimizers of the (asymptotic) absolute relative bias. The key is to notice that

$$A_\psi - A_{theo} = \psi^2 \left([1-\kappa] \sum_{k \in S_r} \widehat{\Gamma}_k \right) + \psi 2 (1-\kappa) n_m + \sum_{k \in S_m} \widehat{h}_{kk} - 2n_m + \sum_{k \in S_r} \Gamma_k^2 := A\psi^2 + B\psi + C,$$

with

$$A := (1-\kappa) \sum_{k \in S_r} \widehat{\Gamma}_k^2,$$

$$B := 2 (1-\kappa) n_m,$$

$$C := \sum_{k \in S_m} \widehat{h}_{kk} - 2n_m - \sum_{k \in S_r} \widehat{\Gamma}_k^2,$$

is a degree two polynomial with positive discriminant $\Delta := B^2 - 4AC$, thus leading to two roots

$$\psi_1 := \frac{-B - \sqrt{\Delta}}{2A}, \qquad \text{and,} \qquad \psi_2 := \frac{-B + \sqrt{\Delta}}{2A},$$

canceling the asymptotic absolute bias.