Catalogue no. 12-001-X
ISSN 1492-0921

**Survey Methodology**


# Generalized regression estimation under misspecified sample design

by Joseph D. Engmark and Jean D. Opsomer

Release date: December 23, 2025

Statistics Canada Statistique Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

**Email at** infostats@statcan.gc.ca

**Telephone,** from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

| | |
|---|---|
| • Statistical Information Service | 1-800-263-1136 |
| • National telecommunications device for the hearing impaired | 1-800-363-7629 |
| • Fax line | 1-514-283-9350 |

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada



**An HTML version is also available.**

*Cette publication est aussi disponible en français*.

Survey Methodology, December 2025
Vol. 51, No. 2, pp. 475-491
Statistics Canada, Catalogue No. 12-001-X


# Generalized regression estimation under misspecified sample design

**Joseph D. Engmark and Jean D. Opsomer**[1]


## Abstract

Classical design-based survey estimation relies on a properly specified sampling design for valid inference. We consider the properties of regression estimation under a misspecified sample design, in which the nominal and true inclusion probabilities do not necessarily match. This general misspecified sample design setting encompasses many challenges in the modern survey environment. Under this setting, an asymptotic analysis of the regression estimator, an expression of the bias, and an expression of the variance are presented. Further, a consistent variance estimator is derived and an expression which estimates the bias in-part or in-whole is discussed. This later expression may be used as an indicator of the presence of bias due to misspecification by a practitioner. A simulation study is conducted to support the presented theory.


**Key Words:** Design-based; Model-assisted; Probability sampling; Survey asymptotics; Survey sampling.

# 1. Introduction

The design-based sampling framework is a robust and general framework for inference on finite population characteristics such as totals or means. Design-based estimators rely on a properly specified sampling design, in which each population element has a known and non-zero probability of selection, for valid inference. However, in the modern survey context this is often no longer a reasonable expectation and the motivation of this paper.

Historically, this problem has been considered in the context of non-response and addressed with non-response adjustments via weighting (Bethlehem, 1988; Särndal and Lundström, 2005). This approach typically incorporates a response mechanism into its estimators, which results in either an unknown response probability or an assumed model for the response. The context considered here is more general: the true design is misspecified and unknown. Unlike in the non-response setting where inference is considered under a joint sampling design and response model, we will assume here that the sample continues to be obtained through a single selection mechanism.

The type of misspecification we will consider occurs when the probability that an element appears in the sample differs from the probability assigned by the nominal sampling design. One major source is when selection processes are not implemented as planned, such as defects in the sampling frame, incorrect or out-of-date measures of size used in probability proportional to size sampling (PPS), incorrect stratification or design variables, not following specified selection methods in the field, incorrect primary sampling unit boundaries, or use of random-route sampling. Misspecification may also be caused by the respondent's effect on the selection processes, such as unaccounted-for non-constant selection probabilities and even non-response (e.g. some individuals are more likely to respond, some animals are more likely to be caught,

1. Joseph D. Engmark, Department of Mathematics, University of Maryland, College Park, MD 20742, USA. E-mail: engmark@umd.edu; Jean D. Opsomer, Department of Mathematics, University of Maryland, College Park, MD 20742, USA. E-mail: jopsomer@umd.edu.

etc.) when those are not treated as an additional phase of sampling. This can also be applied to the general setting of non-probability sampling, when using a (nominal) pseudo-design reflecting how elements were selected to appear in the sample. See Elliott and Valliant (2017) for detailed discussion of inference for non-probability samples.

The generalized regression estimator (GREG) is a well-known model-assisted estimator for a population total. When the sampling design is correctly specified, the regression estimator is asymptotically unbiased and mean square consistent regardless of the quality of the regression model. Moreover, an expression for the asymptotic variance of the estimator is available and the expression can be estimated. These results are presented in Chapters 6.4-6.6, pages 225-238, of Särndal, Swensson and Wretman (1992) (SSW henceforth), a comprehensive text on historical model-assisted estimators, and more recently in Sections 1-6 of Breidt and Opsomer (2017) (BO17 henceforth), a paper presenting an asymptotic framework for model-assisted estimators with the GREG as one specific prediction technique.

The GREG has been studied in the presence of non-response. A key challenge is obtaining expressions and estimators for the bias and variance to be used for inference. Bethlehem (1988) gives an expression of the bias and variance and Särndal and Lundström (2005) give an estimator for the variance. In our more general setting, we produce both an expression for the bias and variance, an estimator for the variance, and an estimator for a portion of the bias when grouping data is available. The bias is the more challenging component to consider. Current sampling literature has discussed the relationship between bias and the correlation between the selection process and the response variable of interest, as well as the need for informative auxiliary information on the selection process to reduce the bias (Meng, 2022; Chambers, Ranjbar, Salvati and Pacini, 2022). The expression for the bias of the GREG presented in this paper affirms this relationship in our context and simulations show the utility of the auxiliary information in reducing the correlation between the selection process and the residual, and hence the bias of the GREG.

The rest of the paper is organized as follows. Section 2 introduces the notation and asymptotic framework of design-based estimators; Section 3 briefly reviews the well-established properties of the GREG under properly specified design and derives the properties under a misspecified design. This section culminates with a feasible variance estimator of the GREG under a misspecified design and an expression which estimates either a portion of the bias or the entire bias, when categorical auxiliary variables are available. In practice, this expression may serve as an indicator of bias or design misspecification. In Section 4, the theory and results are confirmed via simulation. Finally, in Section 5 we conclude the discussion.

# 2. Notation and asymptotic framework

## 2.1 Notation

The framework of this paper is the classical design-based framework, avoiding model assumptions on the variables collected by the survey, as in Section 1 of BO17 and Chapter 2 of SSW. Let $y_k$ denote the value of a variable of interest for element $k$ in population $U = \{1, 2, \ldots, N\}$ and $t_y = \sum_{k \in U} y_k$ the population total, here treated as fixed but unknown prior to sampling.

The *sampling design*, denoted $p(s)$, is a probability distribution on the set of all $2^N$ possible subsets of $U$. That is, $p(s)$ is the probability of selecting the particular sample $s$. Define the sample membership indicator $I_k = 1$ if $k \in s$ and $I_k = 0$ otherwise. For $k, l \in U$, denote the *first-order inclusion probabilities* of the design as $\pi_k = \mathbb{E}[I_k] = P[k \in s] = \sum_{s \subset U:\, k \in s} p(s)$, and the *second-order inclusion probabilities* of the design as $\pi_{kl} = \mathbb{E}[I_k I_l] = P[k, l \in s] = \sum_{s \subset U:\, k, l \in s} p(s)$. The design $p(s)$ is a *probability sampling design* if $\pi_k > 0$ for all $k \in U$ and is said to be *measurable* if $\pi_{kl} > 0$ for all $k, l \in U$.

The Horvitz-Thompson (HT) estimator for a total is given by,

$$\hat{t}_{y\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k.$$

The HT estimator is known to be design unbiased for $t_y$, with variance

$$\operatorname{Var}(\hat{t}_{y\pi}) = \sum_{k, l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Provided the design is measurable, an unbiased estimator of the variance is given by,

$$\widehat{\operatorname{Var}}(\hat{t}_{y\pi}) = \sum_{k, l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

## 2.2 Asymptotic framework

Design-based estimators are constructed from a probability sample selected from the population. It is natural to begin asymptotic arguments from a sequence of finite populations $U_N$ of size $N$, with $N \to \infty$. A sequence of sampling designs $p_N(\cdot)$ is associated with this sequence. For each $N$ in the sequence, a sample $s_N \subset U_N$ is drawn according to design $p_N(\cdot)$, with sample size $n_N$. Inclusion probabilities $\pi_{kN}$, $\pi_{klN}$ are associated with design $p_N(\cdot)$. The subscript $N$ will be dropped wherever possible, as is common practice.

Regularity condition are needed on both the population, $U_N$, and the design, $p_N(\cdot)$, to ensure asymptotic results are well-defined because both change with $N$. The initial discussion is restricted to the simple HT estimator. Section 2 of BO17 introduces the difference estimator, which can be viewed as a shifted HT estimator, followed by the GREG, which can be viewed as a difference estimator using estimated coefficients and a remainder term. Therefore, the asymptotic results of the HT estimator directly carry over to the difference estimator, which in turn carry over to the GREG, provided the remainder term is negligible as compared to the estimated totals. The following regularity conditions are assumed, for a design of non-random size $n$:

D1. As $N \to \infty$, $\frac{n}{N} \to \theta \in (0, 1)$. For all $N$, $\min_{k \in U} \pi_k \geq \lambda_1 > 0$ and $\limsup_{N \to \infty} n \max_{k, l \in U:\, k \neq l} |\Delta_{kl}| < \infty$.

D2. For all $N$, $\min_{k \in U} \pi_{kl} \geq \lambda_2 > 0$.

D3. The study variables $\{y_k\}_{k \in U}$ satisfy $\limsup_{N \to \infty} \frac{\sum_{k \in U} y_k^4}{N} < \infty$.

D4. As $N \to \infty$, $\frac{\hat{t}_{y\pi} - t_y}{\sqrt{\operatorname{Var}(\hat{t}_{y\pi})}} \xrightarrow{d} N(0, 1)$.

D1 and D2 are conditions on the sampling design $p_N(\cdot)$, ensuring selection probabilities are bounded away from zero as necessary for the sample design to be a probability sampling design. D3 is a condition on the sequence of finite populations $U_N$. D1 and D3 are used to show that $\hat{t}_{y\pi}$ is consistent for $t_y$. D2 and D3 are used to show consistent estimation of the variance of the estimator. These conditions combined with D4, the assumed asymptotic normality of the HT estimator, enable the construction of confidence intervals with correct coverage for the HT estimator. In sampling designs with random sample size, the $n$ above denotes the expected sample size, $\mathbb{E}\left(\sum_{k \in U} I_k\right)$. More detailed discussion of the asymptotic framework can be found in BO17, Section 2.

# 3. Regression estimator and properties

## 3.1 The regression estimator (GREG)

We define the GREG and its components. For complete discussion under a properly specified design, see Chapter 6 of SSW and Section 6 of BO17. Let $\mathbf{x}_k$ be a vector of auxiliary variables. Assume population totals $\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$ are known and sample vectors $\{\mathbf{x}_k\}_{k \in s}$ are observed. The well-known calibrated regression estimator (GREG) is defined as:

$$\begin{aligned} \hat{t}_{yr} &= \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \widehat{\mathbf{B}} \\ &= \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \widehat{\mathbf{T}}^{-1} \hat{\mathbf{t}}, \end{aligned} \tag{3.1}$$

where $\hat{\mathbf{t}} = \sum_{k \in s} \mathbf{x}_k y_k / (\pi_k \sigma_k^2)$ and $\widehat{\mathbf{T}} = \sum_{k \in s} \mathbf{x}_k \mathbf{x}_k' / (\pi_k \sigma_k^2)$ are HT estimators for $\mathbf{t} = \sum_{k \in U} \mathbf{x}_k y_k / \sigma_k^2$ and $\mathbf{T} = \sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' / \sigma_k^2$; $\widehat{\mathbf{B}} = \widehat{\mathbf{T}}^{-1} \hat{\mathbf{t}}$ is an estimator for $\mathbf{B} = \mathbf{T}^{-1} \mathbf{t}$; and it is assumed that $\widehat{\mathbf{T}}$ is non-singular. The GREG can also be expressed as $\hat{t}_{yr} = \sum_{k \in s} g_{ks} y_k / \pi_k$, where $g_{ks} = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \widehat{\mathbf{T}}^{-1} \mathbf{x}_k / \sigma_k^2$. For simplicity, we set $\sigma_k^2 = 1$ for all $k \in U$, constraining the scope to homoskedastic multiple regression and omitting the variance parameter from the remainder of the paper.

The GREG is a non-linear function of HT estimators, which can be approximated through *Taylor linearization* as $\hat{t}_{yr,0} = \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \mathbf{B}$ (see SSW Chapter 6.6, page 236). The linearized approximation has variance

$$\operatorname{Var}(\hat{t}_{yr,0}) = \sum_{k,l \in U} \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l},$$

where $E_k = y_k - \mathbf{x}_k' \mathbf{B}$, with estimator

$$\widehat{\operatorname{Var}}(\hat{t}_{yr,0}) = \sum_{s,k,l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_{ks}}{\pi_k} \frac{e_{ls}}{\pi_l},$$

where $e_{ks} = y_k - \mathbf{x}_k' \widehat{\mathbf{B}}$. Under conditions D1-D4, the GREG is asymptotically unbiased and mean square consistent for $t_y$, and the estimator of the approximate variance can be used for asymptotically correct inference.

## 3.2 Regression estimator under misspecified design

The properties above hold assuming the sampling design is properly specified, i.e. $\mathbb{E}(I_k) = \pi_k$, $\mathbb{E}(I_{kl}) = \pi_{kl}$, and $\operatorname{Cov}(I_k, I_l) = \Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. Under a misspecified design, the true first-order inclusion probabilities are $\mathbb{E}(I_k) = \pi_k^* \neq \pi_k$, where $\pi_k$ is the inclusion probability as defined by the nominal sampling design; similarly for the second-order inclusion probability $\mathbb{E}(I_{kl}) = \pi_{kl}^* \neq \pi_{kl}$; and the covariance is $\operatorname{Cov}(I_k, I_l) = \Delta_{kl}^* = \pi_{kl}^* - \pi_k^* \pi_l^* \neq \Delta_{kl}$. Under this setting, HT estimators using the nominal design are no longer unbiased estimators for their corresponding population level quantities. In particular,

$$\mathbb{E}(\hat{t}_{y\pi}) \;=\; \mathbb{E}\left(\sum_{k \in U} I_k \frac{y_k}{\pi_k}\right) \;=\; \sum_{k \in U} \frac{\pi_k^*}{\pi_k} y_k \;\equiv\; t_y^* \;\neq\; t_y.$$

The remaining HT estimators of the GREG components now estimate "*" totals:

$$\mathbb{E}(\hat{\mathbf{t}}_{x\pi}) \;=\; \sum_{k \in U} \frac{\pi_k^*}{\pi_k} \mathbf{x}_k \;\equiv\; \mathbf{t}_x^* \;\neq\; \mathbf{t}_x; \qquad \mathbb{E}(\hat{\mathbf{t}}) \;=\; \sum_{k \in U} \frac{\pi_k^*}{\pi_k} \mathbf{x}_k\, y_k \;\equiv\; \mathbf{t}^* \;\neq\; \mathbf{t};$$

$$\mathbb{E}(\widehat{\mathbf{T}}) \;=\; \sum_{k \in U} \frac{\pi_k^*}{\pi_k} \mathbf{x}_k \mathbf{x}_k' \;\equiv\; \mathbf{T}^* \;\neq\; \mathbf{T}; \qquad \mathbf{T}^{*-1} \mathbf{t}^* \;\equiv\; \mathbf{B}^* \;\neq\; \mathbf{B}.$$

Throughout our discussion, "*" indicates an unknown quantity due to the presences of $\pi_k^*$ in their construction. To support asymptotic statements on HT estimators under a misspecified design, we require an additional regularity condition on the true unknown design:

D5. For all $N$, $\min_{k \in U} \pi_k^* \geq \lambda_1 > 0$ and $\limsup_{N \to \infty} n \max_{k, l \in U : k \neq l} |\Delta_{kl}^*| < \infty$,

ensuring true selection probabilities are bounded away from zero as necessary for the sample design to be a probability sampling design. The asymptotic normality of the HT estimator (D4) is now assumed under the true design. In the case of random sample size, $n$ would now be the expected sample size under the true design.

Through a direct application of Taylor linearization (see SSW Chapter 6.6, page 236), evaluated at the HT estimators new "*" totals, the GREG (3.1) can be expressed as,

$$\hat{t}_{yr} \;=\; \hat{t}_{y\pi} \;+\; (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \mathbf{B}^* \;+\; (\mathbf{t}_x - \mathbf{t}_x^*)' \mathbf{T}^{*-1} (\hat{\mathbf{t}} - \widehat{\mathbf{T}} \mathbf{B}^*) \;+\; O_p\left(\frac{N}{n}\right).$$

As $n \to \infty$, the remainder term is asymptotically negligible compared to the estimated totals, and the GREG continues to behave as a linear combination of HT estimators. The HT estimators are now unbiased and mean square consistent for their respective "*" quantities. Omitting the remainder term, the following expression is a linearized approximation of the GREG under a misspecified design:

$$\tilde{t}_{yr,0} = \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})'\mathbf{B}^* + (\mathbf{t}_x - \mathbf{t}_x^*)'\mathbf{T}^{*-1}(\hat{\mathbf{t}} - \hat{\mathbf{T}}\mathbf{B}^*). \tag{3.2}$$

Expression (3.2), referred to as the *linearized estimator*, is used to investigate the asymptotic properties of the GREG under a misspecified design.

## 3.3 Bias of regression estimator under misspecified design

The approximate bias of the GREG under a misspecified design follows directly from taking the expectation of (3.2) with respect to the design:

$$\mathbb{E}(\tilde{t}_{yr,0}) = t_y^* + (\mathbf{t}_x - \mathbf{t}_x^*)'\mathbf{B}^*,$$

and hence

$$\text{Bias}(\tilde{t}_{yr,0}) = \mathbb{E}(\tilde{t}_{yr,0}) - t_y = \sum_{k \in U} \left( \frac{\pi_k^*}{\pi_k} - 1 \right) E_k^*, \tag{3.3}$$

with $E_k^* = y_k - \mathbf{x}_k'\mathbf{B}^*$.

Expression (3.3) will be referred to as the linearized bias. It implies if $\pi_k^* / \pi_k$ is uncorrelated with $E_k^*$, the bias will be asymptotically negligible. Other representations of the bias can be obtained but will not be pursued here. Here we avoid making explicit modeling assumptions about the data and instead focus on the finite population correlation as the driver of bias. This correlation depends on the predictive power of the auxiliary variables in the survey variable $y_k$ and the ratio of true and nominal inclusion probabilities. We will investigate this empirically in Section 4.

## 3.4 Variance of regression estimator under misspecified design

Variance expressions under a misspecified design are readily obtained using the techniques of SSW (see Chapter 6.6, page 237). The linearized estimator under a misspecified design (3.2) is

$$\begin{aligned} \tilde{t}_{yr,0} &= \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})'\mathbf{B}^* + (\mathbf{t}_x - \mathbf{t}_x^*)'\mathbf{T}^{*-1}(\hat{\mathbf{t}} - \hat{\mathbf{T}}\mathbf{B}^*) \\ &= \mathbf{t}_x'\mathbf{B}^* + \sum_{k \in s} \frac{G_k^* E_k^*}{\pi_k}, \end{aligned}$$

where $G_k^* = 1 + (\mathbf{t}_x - \mathbf{t}_x^*)'\mathbf{T}^{*-1}\mathbf{x}_k$ can be viewed as a population level analog of the sample-dependent g-weights $g_{ks} = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x)'\hat{\mathbf{T}}^{-1}\mathbf{x}_k$. Under the misspecified setting, $g_{ks} \xrightarrow{p} G_k^*$ and $e_{ks} \xrightarrow{p} E_k^*$ with respect to a sequence of increasing sampling designs. Hence, the linearized estimator can again be written as a population fit ($\mathbf{t}_x'\mathbf{B}^*$, a constant) plus the sum of a weighted residual $\left(\sum_{k \in s} G_k^* E_k^* / \pi_k, \text{ an estimator}\right)$, an estimator). As a difference estimator, the variance and a variance estimator are immediately available:

$$\text{Var}(\tilde{t}_{yr,0}) = \sum_{k,l \in U} \Delta_{kl}^* \left( \frac{G_k^* E_k^*}{\pi_k} \right) \left( \frac{G_l^* E_l^*}{\pi_l} \right), \tag{3.4}$$

$$\widetilde{\text{Var}}(\tilde{t}_{yr,0}) = \sum_{k,l \in s} \frac{\Delta_{kl}^*}{\pi_{kl}} \left( \frac{g_{ks} e_{ks}}{\pi_k} \right) \left( \frac{g_{ls} e_{ls}}{\pi_l} \right). \tag{3.5}$$

However, the variance estimator (3.5) is not feasible due to the unknown covariance, $\Delta_{kl}^{*}$. Instead of attempting to study the case of general unknown $\Delta_{kl}^{*}$, we will restrict our attention to the setting, common in large-scale surveys, in which treating the design as being with-replacement is reasonable.

Substituting with-replacement sampling results for without-replacement sampling results is reasonable for large populations, where the probability of selecting an element more than once is vanishingly small. Further, it is known that a with-replacement sampling design generally increases the variance estimate as compared to without-replacement sampling due to the omission of the finite population correction, while the estimated linearized variance tends to underestimate in practice. Taken together, these opposite effects results in a variance estimator with reasonable properties in practice.

The derivation of the with-replacement variance estimator follows from result 2.9.1 of SSW (page 51), where an unbiased variance estimator for the "p-expanded with replacement" estimator is presented. Following this approach, we define $p_k = \pi_k / n$ for $k = 1, \ldots, N$, where $p_k = \mathbb{P}$ (selecting element $k$ on a given draw under the nominal design), ignoring the possibility that an element can appear in the sample more than once. Let $Z_i^{*}, i = 1, \ldots, n$ be random variables such that, $Z_i^{*} = G_k^{*} E_k^{*} / p_k$ if element $k$ is selected on draw *i*. The distribution of $Z_i^{*}$ is $\mathbb{P}\left(Z_i^{*} = G_k^{*} E_k^{*} / p_k\right) = p_k^{*}$, where $p_k^{*}$ is unknown due to misspecification and defined through $\pi_k^{*} = n p_k^{*}$. With a slight abuse of notation, we will suppress the draw subscript $i = 1, \ldots, n$, and write $Z^{*}$ to denote the random variable and $k \in s$ to indicate the elements in the sample. Then,

$$\operatorname{Var}\left(\sum_{k \in s} \frac{G_k^{*} E_k^{*}}{\pi_k}\right) = \operatorname{Var}\left(\bar{Z}^{*}\right) \tag{3.6}$$

with $\bar{Z}^{*} = \sum_{k \in s} Z_k^{*} / n$. We note that although (3.6) and (3.4) are of the same form, they are evaluated under different sampling distributions. Under with-replacement sampling, assumptions on the covariance elements of D1 and D4 are trivially satisfied because off-diagonal elements of the covariance matrix are equal to zero. Similarly, regularity condition D2 is no longer necessary because $\pi_{kl} = n^2 \pi_k \pi_l > 0$ by conditions on $\pi_k$ in D1.

The $Z^{*}$'s are drawn independently on the same set of probabilities $p_1^{*}, \ldots, p_N^{*}$ in each draw. Since the $Z^{*}$'s are i.i.d., $\operatorname{Var}(\bar{Z}^{*}) = n^{-1} \operatorname{Var}(Z^{*})$ where, $\operatorname{Var}(Z^{*}) = \sum_{k \in U}\left(Z_k^{*} - \mathbb{E}(Z^{*})\right)^2 p_k^{*}$. An unbiased variance estimator of $\operatorname{Var}(Z^{*})$ is $\widetilde{\operatorname{Var}}(Z^{*}) = (n-1)^{-1} \sum_{k \in s}\left(Z_k^{*} - \bar{Z}^{*}\right)^2$. This contains unknown "*" quantities, motivating the feasible estimator $\widehat{\operatorname{Var}}(Z^{*}) = (n-1)^{-1} \sum_{k \in s}\left(z_{ks} - \bar{z}_s\right)^2$, where $z_{ks} = g_{ks} e_{ks} / p_k$ and $\bar{z}_s = n^{-1} \sum_{k \in s} z_{ks}$. Putting these components together, the feasible variance estimator for the GREG under misspecified design is

$$\widehat{\operatorname{Var}}(\tilde{t}_{yr,0}) = \frac{1}{n} \widehat{\operatorname{Var}}(Z^{*}) = \frac{1}{n(n-1)} \sum_{k \in s}\left(z_{ks} - \bar{z}_s\right)^2. \tag{3.7}$$

In Section S.1 of the supplementary material (see Engmark and Opsomer, 2025) it is shown the feasible variance estimator (3.7) is consistent for the variance of the linearized estimator (3.4), under the with-replacement sampling distribution. Therefore in practice, misspecification of the design does not affect the ability to consistently estimate the variance of the regression estimator under with-replacement sampling.

We postulate that this result continues to hold as long as the correlation structure of the nominal and misspecified design are the same, i.e. $\pi_{kl} / \pi_k \pi_l = \pi_{kl}^* / \pi_k^* \pi_l^*$ (which is trivially true under with-replacement sampling). In practice, this would mean the stratification and clustering of the design remain correctly specified. We do not pursue this generalization formally here and focus on use of (3.7) for variance estimation.

## 3.5 Controlling bias of regression estimator under misspecified design

Directly estimating the approximate bias expression (3.3) is not feasible due to the unknown $\pi_k^*$. Here, we propose a measure that can be useful to detect bias in a common setting of social surveys, in which most or all auxiliary variables are categorical. In this case, the potential impact of the auxiliary variables on the bias would only occur through the group structure in the unknown selection probabilities determined by the set of values of these variables, but it is not known which of the variables determine that structure. We consider a scenario in which we are trying to assess whether a particular auxiliary variable is informative on the selection probabilities, in the sense that omitting it from the predictors in the GREG would cause bias. If this auxiliary variable is indeed determined to be informative, it can then be added to the GREG predictors to reduce the bias caused by the misspecified design.

The values of the categorical variable under consideration partition the population into a fixed number of groups, $g = 1, \ldots, G,$ where we require $G \leq |s| = n.$ Define $\theta_g$ as the average selection probability for group $g$ in the population, $\theta_g = N_g^{-1} \sum_{k \in U_g} \pi_k^*.$ The linearized bias (3.3) can be expressed as

$$\begin{aligned}\operatorname{Bias}(\tilde{t}_{yr,0}) &= \sum_{g=1}^{G} \sum_{k \in U_g} \left( \frac{\pi_k^*}{\pi_k} - 1 + \frac{\theta_g}{\pi_k} - \frac{\theta_g}{\pi_k} \right) E_k^* \\ &= \sum_{g=1}^{G} \sum_{k \in U_g} \frac{1}{\pi_k} \left( \theta_g - \pi_k \right) E_k^* + \sum_{g=1}^{G} \sum_{k \in U_g} \frac{1}{\pi_k} \left( \pi_k^* - \theta_g \right) E_k^* \\ &= B_1 + B_2.\end{aligned}$$

Loosely speaking, $B1$ can be viewed as a “between-group bias” and $B2$ as a “within-group bias” of (3.2).

We first consider the case in which the selection probabilities actually match the group structure, i.e. $\pi_k^* = \theta_g$ for all $k \in U_g,$ $g = 1, \ldots, G.$ In this case, the bias component $B_2$ will be 0. Because the quantity $N_g^0 = \sum_{k \in U_g} 1 / \pi_k$ is known from the nominal sampling design and $\widehat{N}_g = \sum_{k \in s_g} 1 / \pi_k$ is the HT estimator of

$$N_g^* = \sum_{k \in U_g} \frac{\pi_k^*}{\pi_k} = \theta_g N_g^0,$$

$\theta_g$ can be estimated by $\hat{\theta}_g = \hat{N}_g / N_g^0 = \left( \sum_{k \in s_g} 1 / \pi_k \right) \Big/ \left( \sum_{k \in U_g} 1 / \pi_k \right).$ Hence, we can define a plug-in estimator of $B_1$ by

$$\widehat{B}_1 = \sum_{g=1}^{G} \sum_{k \in s_g} \frac{1}{\hat{\theta}_g} \frac{1}{\pi_k} \left( \hat{\theta}_g - \pi_k \right) e_{ks}. \tag{3.8}$$

Hence, if this group propensity structure holds, $B_2$ will be exactly zero and $\widehat{B}_1$ will be approximately unbiased for $B_1$ and estimate the linearized bias in its entirety.

If the group structure does not hold in the selection probabilities, $B_2$ will be non-zero and $\widehat{B}_1$ will be a biased estimator of $B_1$. A large value of $\widehat{B}_1$ may nevertheless serve as a potential indicator of bias due to design misspecification. Section S.2 of the supplementary material show the behavior of $\widehat{B}_1$ starting from the general setting in which the selection probabilities are not assumed to be constant with the groups.

# 4. Simulation

The goal of the simulation study is to assess the bias of the GREG under the misspecified design, the ability of the variance estimator to recover the design variance and of the bias measure to indicate the need to add a calibration variable.

## 4.1 Simulation setup

### 4.1.1 Data generation

The simulated population consists of $\{y_k, x_{1k}, x_{2k}, w_k, z_k, \pi_k, \pi_k^*\}$ for $k = 1, \ldots, N.$ The variables of interest in the populations are created through the mean functions $\mu_i(\mathbf{x}), i = 1, \ldots, 6$ given below. The functions follow from Section 3 of Breidt and Opsomer (2000).

linear: $\mu_1(x_1, x_2) = 1 + 3x_1 + 5x_2 + \varepsilon,$

quadratic: $\mu_2(x_1, x_2) = 1 + 3x_1 + 4(x_2 - 4)^2 + \varepsilon,$

bump: $\mu_3(x_1, x_2) = 1 + 3x_1 + 10x_2 + 25\exp(-2(x_2 - 4)^2) + \varepsilon,$

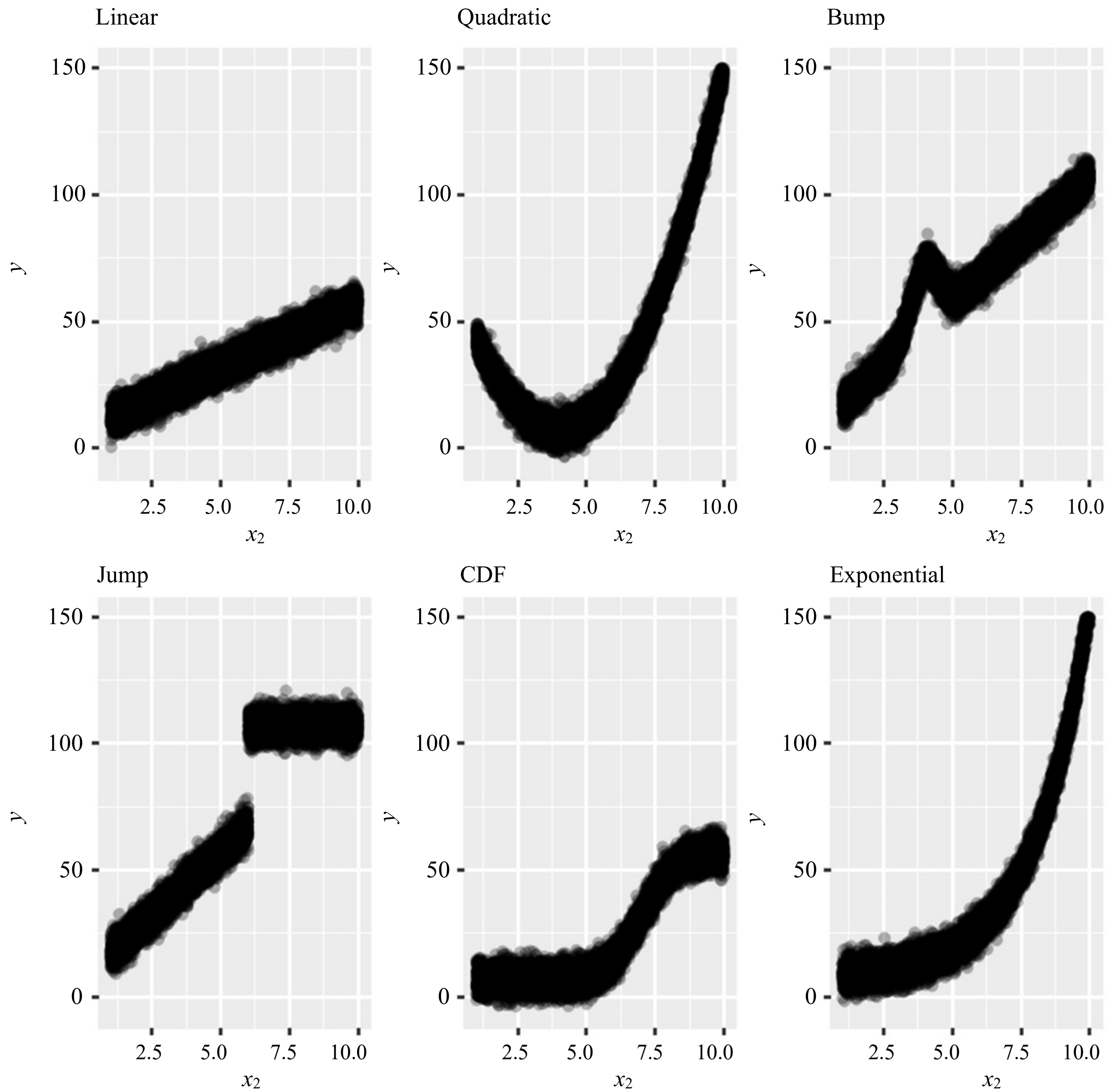
jump: $\mu_4(x_1, x_2) = 1 + 3x_1 + 10x_2 I_{\{x_2 \le 6\}} + 100 I_{\{x_2 > 6\}} + \varepsilon,$

cdf: $\mu_5(x_1, x_2) = 1 + 3x_1 + 50\Phi\left(x_2 - 7\right) + \varepsilon,$ with $\Phi$ the standard normal cdf,

exponential: $\mu_6(x_1, x_2) = 1 + 3x_1 + \exp\left((1/2)\, x_2\right) + \varepsilon.$

The mean functions represent a range of different relationships between $y$ and $(x_1, x_2),$ as illustrated in Figure 4.1 and are used to demonstrate the behavior of the GREG in a typical survey context, in which the same weights are applied to all the survey variables. For $\mu_1,$ the data are linear in $x_2,$ and the GREG is expected to perform well. For the other mean functions, it is of interest to see how the GREG performs as the data depart from linear: for $\mu_2,$ the data are quadratic in $x_2;$ $\mu_3,$ the data has a local departure from linear in $x_2$ in the form of a "bump"; $\mu_4,$ the data has a discontinuity in $x_2;$ $\mu_5,$ the data are sigmoidal in $x_2;$ $\mu_6,$ the data are exponential in $x_2.$

**Figure 4.1 Plot of mean functions: linear, quadratic, bump, jump, cdf, exponential**




**Note:** CDF = cumulative distribution function.

The population auxiliary variables $x_{1k}$ and $x_{2k}$ are generated as i.i.d. uniform (1, 3) random variables and uniform (1, 10) random variables, respectively, and are independent of one another. The true design in the simulation is dependent on $x_2$. The population variables of interest, $y_k$, are generated from the mean functions by adding i.i.d. $N(0, \sigma^2)$ errors, with $\sigma = 3$. The population size is $N =$ 10,000 and sample size is $n = 99$.

### 4.1.2 Sample selection

The aim is to simulate settings where there is bias in the regression estimator. The sample is selected such that the design is informative, as controlled by the auxiliary variable $x_2$, i.e. both the variable of interest $y_k$ and the true selection probabilities depend on $x_2$. The nominal design is not dependent on $x_2$. Four different settings are considered and they will be labeled Setting I - Setting IV.

In Setting I, the nominal design is simple random sampling without replacement (SI), and the true design is probability proportional to size sampling (PPS) with respect to $x_2$.

In Setting II, we simulate non-constant inclusion probabilities in both the nominal and the true design. Setting II is a stratified version of Setting I, the nominal design is stratified simple random sampling without replacement (STSI), and the true design is stratified probability proportional to size sampling (STPPS) with respect to $x_2$. The strata are determined by a variable $w_k$, which is correlated with $y_k$ and constructed by adding $N(0, 1)$ errors to $x_{1k}$. The population is sorted by $w_k$, and three strata are formed with sizes $N_1$ = 5,000, $N_2$ = 3,000, and $N_3$ = 2,000, respectively, based on the values of $w_k$. From each stratum, samples of size $n_1 = n_2 = n_3 = 33$ are drawn.

In Setting III and Setting IV, we investigate the behavior of the bias component estimator when the selection probabilities are constant or non-constant within groups. In these settings, the population elements are split into both groups and strata. The groups are determined by a variable $z_k$, which is correlated with $y_k$ and constructed by adding $N(0, 1)$ errors to $x_{2k}$. The population is sorted by $z_k$, and two groups of equal size are formed, based on the values of $z_k$ and without regard for the strata. The strata are determined as above. In Setting III the nominal design is STSI, and the true design is STSI with constant group inclusion probabilities within strata. In Setting IV the nominal design is STSI, and the true design is STPPS with respect to $x_2$, with non-constant group inclusion probabilities within strata. A more detailed description of sampling process is described in Section S.3 of the supplementary material.

### 4.1.3 Estimators

The Horvitz-Thompson estimator under the nominal design, denoted $\mathrm{HT}(\pi)$, serves as the naive estimator by only using the nominal design in its construction. The GREG under the nominal design calibrated on $\mathbf{x}_1$ alone, $\mathbf{x}_2$ alone, and both $\mathbf{x}_1$ and $\mathbf{x}_2$, denoted $\mathrm{GREG}(\mathbf{x}_1, \pi)$, $\mathrm{GREG}(\mathbf{x}_2, \pi)$, $\mathrm{GREG}(\mathbf{x}_1\mathbf{x}_2, \pi)$ respectively, serve as practical estimators when a design is misspecified and auxiliary information is available. For a full comparison the aforementioned estimators are also calculated under the true unknown design, denoted $\mathrm{HT}(\pi^*)$, $\mathrm{GREG}(\mathbf{x}_1, \pi^*)$, $\mathrm{GREG}(\mathbf{x}_2, \pi^*)$, $\mathrm{GREG}(\mathbf{x}_1\mathbf{x}_2, \pi^*)$, and act as ideal estimators.

For more direct comparison of variance estimation, the variance of the $\mathrm{HT}(\pi)$ is also estimated by the "p-expanded with-replacement" variance estimator (equation 2.9.9, page 52 of SSW), under the assumption each element in the sample only appears once, as used in the derivation of (3.7):

$$\widehat{\mathrm{Var}}(\hat{t}_{y\pi}) = \frac{1}{n(n-1)} \sum_{k \in s} \left( n \frac{y_k}{\pi_k} - \hat{t}_{y\pi} \right)^2. \tag{4.1}$$

When the nominal design is STSI, stratification is accounted for in the variance estimation of $\mathrm{HT}(\pi)$ by estimating the variance within each sample strata via (4.1) and summing over all strata.

The variances for the GREG under the nominal design are estimated by (3.7). When the nominal design is STSI, stratification is accounted for in the variance estimation of the GREG by estimating the variance within each sample strata via (3.7) and summing over all strata

$$\widehat{\mathrm{Var}}_{\mathrm{ST}}(\hat{t}_{yr}) = \sum_{h=1}^{H} \frac{1}{n_h(n_h - 1)} \sum_{k \in s_h} \left(z_{hks} - \bar{z}_{hs}\right)^2,$$

where $z_{hks} = g_{ks} e_{ks} / p_{hk}$, $\bar{z}_{hs} = n_h^{-1} \sum_{k \in s_h} z_{hks}$, $p_{hk} = \pi_k / n_h$, with $e_{ks}$ and $g_{ks}$ as defined in Sections 3.1 and 3.4, respectively.

The variance estimation of $\mathrm{HT}(\pi^*)$ and the GREG under the true design follow by replacing $\pi_k$ with $\pi_k^*$ in the estimators previously described. Again, when the true design is STPPS, stratification is accounted for when estimating the variance of $\mathrm{HT}(\pi^*)$ and the GREG.

## 4.2 Simulation results

The population is kept fixed during 10,000 replicate samples allowing the design-averaged performance of the estimators to be evaluated. For each sample in the simulation, the estimated total and estimated variance of each estimator, generically referred to here as $\hat{t}$ and $\widehat{\mathrm{Var}}(\hat{t})$, are calculated. Finally, a confidence interval (CI), $\hat{t} \pm z_{1-\alpha/2} \widehat{\mathrm{Var}}(\hat{t})^{1/2}$, is constructed with $z_{1-\alpha/2}$ for $\alpha = 0.05$.

These values combined with the known population total, $t_y$, allow us to calculate the following elements found in the tables below: (i) Relative bias of the estimators, $(\mathbb{E}(\hat{t}) - t_y) / t_y$; (ii) Bias ratio of the estimators, $\mathrm{BR}(\hat{t}) = (\mathbb{E}(\hat{t}) - t_y) / \mathrm{Var}(\hat{t})^{1/2}$; (iii) Mean width of the confidence intervals, normalized by the mean width of the confidence intervals under $\mathrm{HT}(\pi^*)$; (iv) Coverage probability of the confidence intervals; (v) Relative bias of the variance estimates, $\left(\mathbb{E}(\widehat{\mathrm{Var}}(\hat{t})) - \mathrm{Var}(\hat{t})\right) / \mathrm{Var}(\hat{t})$.

We also report the correlation between population level residuals and the ratio $\pi_k^* / \pi_k$ for $k = 1, \ldots, N$, $\mathrm{Corr}\left(E_k^*, \pi_k^* / \pi_k\right)$. This value is not based on simulation, rather a characteristics of the population and the driver of the bias we see in the estimators. For $\mathrm{HT}(\pi)$ we report $\mathrm{Corr}\left(y_k, \pi_k^* / \pi_k\right)$. For the estimators under the true design, we omit the correlations which are trivially equal to zero. The simulation results for the first three survey variables are shown in Tables 4.1 and 4.2. The complete results including the remaining survey variables are presented in Table S.1 and Table S.2 of the supplementary material.

Tables 4.1 and 4.2 show the GREG calibrated on the non-informative variable alone, $\mathbf{x}_1$, behaves similarly to the naive HT estimator, $\mathrm{HT}(\pi)$. Both have significant bias driven by the correlation, similarly wide CIs, very poor coverage, and negligible bias in the variance estimator. When the GREG is calibrated on the informative variable, $\mathbf{x}_2$ or $\mathbf{x}_1\mathbf{x}_2$, the bias and correlation are reduced, CIs are smaller as compared to the naive HT estimator, coverage is improved as compared to the naive HT estimator, and the variance estimator has negligible bias. In many case the bias of the variance estimator used for the nominal design is smaller than that for the true design, yet with no significant difference in magnitude. Poor coverage occurs because of the bias in the estimators not the variance estimators.

**Table 4.1**
**Simulation results under Setting I – *Nominal design - SI, true design - PPS w.r.t.* $x_2$**

| Population | Estimator | Relative Bias | $\text{Corr}\left(E_k^*, \frac{\pi_k^*}{\pi_k}\right)$ | Normalized CI Width | Coverage | Relative Bias of Variance Est. |
|---|---|---|---|---|---|---|
| linear | HT $(\pi)$ | 0.18 | 0.97 | 2.05 | 0.00 | -0.01 |
| | GREG $(\mathbf{x}_1, \pi)$ | 0.18 | 0.97 | 2.03 | 0.00 | -0.02 |
| | GREG $(\mathbf{x}_2, \pi)$ | 0.00 | 0.01 | 0.67 | 0.94 | -0.02 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi)$ | 0.00 | 0.02 | 0.58 | 0.94 | -0.02 |
| | HT $(\pi^*)$ | 0.00 | | – | 0.93 | 0.01 |
| | GREG $(\mathbf{x}_1, \pi^*)$ | 0.01 | | 2.93 | 0.92 | -0.08 |
| | GREG $(\mathbf{x}_2, \pi^*)$ | 0.00 | | 0.68 | 0.93 | -0.06 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi^*)$ | 0.00 | | 0.58 | 0.93 | -0.07 |
| quadratic | HT $(\pi)$ | 0.35 | 0.79 | 1.63 | 0.08 | -0.01 |
| | GREG $(\mathbf{x}_1, \pi)$ | 0.35 | 0.79 | 1.63 | 0.08 | -0.02 |
| | GREG $(\mathbf{x}_2, \pi)$ | -0.15 | -0.47 | 1.16 | 0.49 | -0.07 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi)$ | -0.15 | -0.47 | 1.15 | 0.48 | -0.09 |
| | HT $(\pi^*)$ | 0.00 | | – | 0.94 | -0.01 |
| | GREG $(\mathbf{x}_1, \pi^*)$ | 0.01 | | 1.36 | 0.95 | -0.02 |
| | GREG $(\mathbf{x}_2, \pi^*)$ | -0.01 | | 1.16 | 0.91 | -0.15 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi^*)$ | -0.02 | | 1.13 | 0.90 | -0.19 |
| bump | HT $(\pi)$ | 0.18 | 0.95 | 1.89 | 0.00 | -0.02 |
| | GREG $(\mathbf{x}_1, \pi)$ | 0.18 | 0.96 | 1.88 | 0.00 | -0.03 |
| | GREG $(\mathbf{x}_2, \pi)$ | 0.01 | 0.18 | 0.88 | 0.91 | -0.04 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi)$ | 0.01 | 0.18 | 0.86 | 0.90 | -0.05 |
| | HT $(\pi^*)$ | 0.00 | | – | 0.94 | 0.01 |
| | GREG $(\mathbf{x}_1, \pi^*)$ | 0.01 | | 3.07 | 0.91 | -0.10 |
| | GREG $(\mathbf{x}_2, \pi^*)$ | 0.00 | | 0.86 | 0.93 | -0.07 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi^*)$ | 0.00 | | 0.84 | 0.93 | -0.09 |

**Note:** CI = Confidence Interval; GREG = Regression estimator; HT = Horvitz-Thompson; PPS = probability proportional to size sampling; SI = simple random sampling without replacement; w.r.t. = with respect to.

Inference under the nominal design is significantly improved under all population data types when calibrating on the informative variable. If the data are linear or slightly depart from linear (bump), the coverage and inference are very good. If the data are non-linear (quadratic), the coverage and inference is significantly improved compared to the naive HT estimator.

**Table 4.2**
**Simulation results under Setting II – *Nominal design - STSI, true design - STPPS w.r.t.* $x_2$**

| Population | Estimator | Relative Bias | Corr$\left(E_k^*, \frac{\pi_k^*}{\pi_k}\right)$ | Normalized CI Width | Coverage | Relative Bias of Variance Est. |
|---|---|---|---|---|---|---|
| linear | HT $(\pi)$ | 0.18 | 0.97 | 2.11 | 0.00 | 0.00 |
| | GREG $(\mathbf{x}_1, \pi)$ | 0.18 | 0.97 | 2.08 | 0.00 | -0.01 |
| | GREG $(\mathbf{x}_2, \pi)$ | 0.00 | 0.01 | 0.68 | 0.94 | -0.01 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi)$ | 0.00 | 0.02 | 0.59 | 0.94 | -0.05 |
| | HT $(\pi^*)$ | 0.00 | | – | 0.93 | 0.02 |
| | GREG $(\mathbf{x}_1, \pi^*)$ | 0.01 | | 2.97 | 0.92 | -0.07 |
| | GREG $(\mathbf{x}_2, \pi^*)$ | 0.00 | | 0.68 | 0.93 | -0.05 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi^*)$ | 0.00 | | 0.59 | 0.93 | -0.10 |
| quadratic | HT $(\pi)$ | 0.34 | 0.79 | 1.63 | 0.12 | 0.00 |
| | GREG $(\mathbf{x}_1, \pi)$ | 0.34 | 0.79 | 1.62 | 0.11 | -0.01 |
| | GREG $(\mathbf{x}_2, \pi)$ | -0.15 | -0.47 | 1.15 | 0.51 | -0.01 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi)$ | -0.15 | -0.47 | 1.14 | 0.51 | -0.04 |
| | HT $(\pi^*)$ | 0.00 | | – | 0.95 | 0.03 |
| | GREG $(\mathbf{x}_1, \pi^*)$ | 0.01 | | 1.36 | 0.95 | -0.01 |
| | GREG $(\mathbf{x}_2, \pi^*)$ | -0.02 | | 1.13 | 0.91 | -0.11 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi^*)$ | -0.02 | | 1.10 | 0.90 | -0.16 |
| bump | HT $(\pi)$ | 0.18 | 0.95 | 1.90 | 0.00 | 0.00 |
| | GREG $(\mathbf{x}_1, \pi)$ | 0.18 | 0.96 | 1.88 | 0.00 | -0.01 |
| | GREG $(\mathbf{x}_2, \pi)$ | 0.01 | 0.18 | 0.89 | 0.91 | -0.04 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi)$ | 0.01 | 0.18 | 0.86 | 0.90 | -0.06 |
| | HT $(\pi^*)$ | 0.00 | | – | 0.94 | 0.00 |
| | GREG $(\mathbf{x}_1, \pi^*)$ | 0.01 | | 3.03 | 0.91 | -0.07 |
| | GREG $(\mathbf{x}_2, \pi^*)$ | 0.00 | | 0.86 | 0.93 | -0.08 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi^*)$ | 0.00 | | 0.83 | 0.92 | -0.11 |

**Note:** CI = Confidence Interval; GREG = Regression estimator; HT = Horvitz-Thompson; STPPS = stratified probability proportional to size sampling; STSI = stratified simple random sampling without replacement; w.r.t. = with respect to.

Second, consider the estimators under the true design. All estimators are unbiased, the CIs are smaller than those under the ideal HT estimator, $\mathrm{HT}(\pi^*)$, when the data are linear or slightly depart from linear (bump) and the estimator is calibrated on an informative variable. When the data are non-linear (quadratic), we see the CIs of the GREG is not smaller than the ideal HT estimator, regardless of the calibration variables. All estimators have good coverage for generally valid inference. The ideal HT estimator has negligible bias in the variance estimator. The GREG across all calibrations has a mild negative bias in the variance estimator. This is attributed to using the linearized variance estimator which tends to underestimate.

The CI interval width, coverage, and bias of the variance estimator behave slightly more poorly than expected. When the simulation is repeated with a larger population and sample size of $N =$ 100,000 and

$n = 999$, we see the CI interval width is improved over the HT estimator when calibrating on the informative variable, the coverage approaches 95%, and relative bias of the variance estimator is zero.

Next, consider the group settings. Under Setting III and Setting IV, for each replicate sample we calculate two additional estimators for the GREG under the nominal design: the bias component estimator $\widehat{B}_1$ and a feasible relative bias estimator $\widehat{B}_1 / \hat{t}$. These allow us to calculate (vi) Expected value of the feasible relative bias estimator, $\mathbb{E}(\widehat{B}_1 / \hat{t})$; (vii) Expected value of bias component estimate scaled by the true bias, $\mathbb{E}(\widehat{B}_1) / \text{Bias}$.

Additionally, the relative bias of the estimators (i) are presented for comparison against the estimated value of (vi). Value (vii) indicates on average what portion of the overall bias is captured by the bias component estimator. Lastly, we report population characteristic $B_1 / (B_1 + B_2)$, which indicates the proportion of the linearized bias that is between-group bias, $B_1$. Results are presented in Tables 4.3 and 4.4 below.

Table 4.3 summarizes Setting III and the ideal condition of constant group propensity. Here, all the bias is within-group bias. The bias component estimator, $\widehat{B}_1$, estimates the entire bias on average, and the practical relative bias estimator, $\widehat{B}_1 / \hat{t}$, estimates the true relative bias well on average. This estimate is best when the data are linear or slightly depart from linear, but also reasonable when the data are non-linear.

Table 4.4 summarizes Setting IV and the practical conditions of non-constant group propensity. Here, only a portion of the bias is within-group bias. The bias component estimator, $\widehat{B}_1$, estimates a portion of the bias on average, and the practical relative bias estimator, $\widehat{B}_1 / \hat{t}$, no longer accurately estimates the true relative bias well on average. However, the practical relative bias estimator does capture a portion and the direction of the true relative bias, which gives it utility as an indicator of bias.

**Table 4.3**
**Bias estimation simulation results under Setting III – *Nominal design - STSI, true design - STSI w.r.t. strata and groups***

| Population | Estimator | Relative Bias | $\mathbb{E}\left(\frac{\widehat{B}_1}{\hat{t}}\right)$ | $\frac{\mathbb{E}(\widehat{B}_1)}{\textbf{Bias}}$ | $\frac{B_1}{B_1 + B_2}$ |
|---|---|---|---|---|---|
| linear | GREG $(\mathbf{x}_1, \pi)$ | 0.10 | 0.09 | 0.99 | 1.00 |
| | GREG $(\mathbf{x}_2, \pi)$ | 0.00 | 0.00 | -0.99 | 1.00 |
| | GREG $(\mathbf{x}_1 \mathbf{x}_2, \pi)$ | 0.00 | 0.00 | 0.07 | 1.00 |
| quadratic | GREG $(\mathbf{x}_1, \pi)$ | 0.20 | 0.16 | 0.99 | 1.00 |
| | GREG $(\mathbf{x}_2, \pi)$ | -0.04 | -0.04 | 0.93 | 1.00 |
| | GREG $(\mathbf{x}_1 \mathbf{x}_2, \pi)$ | -0.04 | -0.04 | 0.92 | 1.00 |
| bump | GREG $(\mathbf{x}_1, \pi)$ | 0.10 | 0.09 | 0.99 | 1.00 |
| | GREG $(\mathbf{x}_2, \pi)$ | 0.00 | 0.00 | 1.03 | 1.00 |
| | GREG $(\mathbf{x}_1 \mathbf{x}_2, \pi)$ | 0.00 | 0.00 | 1.02 | 1.00 |

**Note:** GREG = Regression estimator; STSI = stratified simple random sampling without replacement; w.r.t. = with respect to.

**Table 4.4**
**Bias estimation simulation results under Setting IV –** ***Nominal design - STSI, true design - STPPS w.r.t. strata and groups***

| Population | Estimator | Relative Bias | $\mathbb{E}\left(\frac{\widehat{B}_1}{\hat{t}}\right)$ | $\frac{\mathbb{E}(\widehat{B}_1)}{\text{Bias}}$ | $\frac{B_1}{B_1+B_2}$ |
|---|---|---|---|---|---|
| linear | GREG $(\mathbf{x}_1, \pi)$ | 0.16 | 0.08 | 0.57 | 0.63 |
| | GREG $(\mathbf{x}_2, \pi)$ | 0.00 | 0.00 | -0.11 | 0.28 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi)$ | 0.00 | 0.00 | -0.16 | 0.33 |
| quadratic | GREG $(\mathbf{x}_1, \pi)$ | 0.31 | 0.18 | 0.76 | 0.65 |
| | GREG $(\mathbf{x}_2, \pi)$ | -0.14 | -0.03 | 0.15 | 0.60 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi)$ | -0.14 | -0.03 | 0.15 | 0.60 |
| bump | GREG $(\mathbf{x}_1, \pi)$ | 0.16 | 0.07 | 0.50 | 0.59 |
| | GREG $(\mathbf{x}_2, \pi)$ | 0.01 | -0.01 | -0.45 | -0.02 |
| | GREG $(\mathbf{x}_1\mathbf{x}_2, \pi)$ | 0.01 | 0.00 | -0.45 | -0.02 |

**Note:** GREG = Regression estimator; STPPS = stratified probability proportional to size sampling; STSI = stratified simple random sampling without replacement; w.r.t. = with respect to.

# 5. Conclusion

A misspecified design introduces bias into classical survey estimators, as is well known. For the HT estimator, the bias is driven by the correlation between the variable of interest and the discrepancy between the nominal and true inclusion probability. For the GREG, the bias is driven by the correlation between the residuals and the same discrepancy.

The GREG can reduce this correlation and correct the bias when the regression fits the data well and the correlation between the response mechanism and the residuals is small. That is, the GREG performs best when the data are close to linear and the regression variables are informative of the true inclusion probabilities. When the data are non-linear, the GREG can reduce the correlation and improve the bias due to misspecification provided the regression variables are informative. This correction significantly improves inference and shows the utility of the GREG regardless of the shape of the data. Thus, the GREG is robust to either misspecification of the selection process or the implied regression model. If the motivating regression model is correct, then incorrect selection probabilities do not cause bias. If the selection probabilities are correct, then an incorrect motivating regression model does not cause bias. In that sense the GREG is "doubly robust".

This extends the literature on the application of the GREG from specific design departures such as non-response, to a general misspecified framework. Further, under this general framework, we have proposed a variance estimator that is a reasonable estimator under all data and design settings and regardless of the regression variables used.

Finally, suppose we have an auxiliary variable which we can form groups from and is believed to be related to the response mechanisms. If there is a constant selection probability within groups, we can estimate the bias of the GREG. If there is an informative, but non-constant group propensity, we can estimate a portion of the bias of the GREG. Such an estimate may be used as an indicator of bias or misspecification by practitioners.

# References

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Offcial Statistics*, 4(3), 251-260.

Breidt, F.J. and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4), 1026-1053.

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

Chambers, R., Ranjbar, S., Salvati, N. and Pacini, B. (2022). Weighting, informativeness and causal inference, with an application to rainfall enhancement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185.

Elliott, M.R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Engmark, J.D. and Opsomer, J.D. (2025). Supplementary material: Generalized regression estimation under misspecified sample design. Available at: https://doi.org/10.13140/RG.2.2.29212.19848.

Meng, X.-L. (2022). Comments on "Statistical inference with non-probability survey samples" – Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. *Survey Methodology*, 48(2), 339-360. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00006-eng.pdf.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.