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

Survey Methodology

# Survey Methodology
# 46-2

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

Release date: December 15, 2020

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** STATCAN.infostats-infostats.STATCAN@canada.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 |

**Depository Services Program**

| | |
|---|---|
| • Inquiries line | 1-800-635-7943 |
| • Fax line | 1-800-565-7757 |

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. 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

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

December 2020

•

Volume 46

•

Number 2

Statistics Canada Statistique Canada

Canada

# SURVEY METHODOLOGY

## A Journal Published by Statistics Canada

*Survey Methodology* is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

# EDITORIAL POLICY

*Survey Methodology* publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

## Submission of Manuscripts

*Survey Methodology* is published twice a year in electronic format. Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/surveymethodology). To communicate with the Editor, please use the following email: (statcan.smj-rte.statcan@canada.ca).

**Survey Methodology**
A Journal Published by Statistics Canada
Volume 46, Number 2, December 2020

## Contents

**Regular Papers**

Survey Methodology, December 2020
Vol. 46, No. 2, pp. 145-180
Statistics Canada, Catalogue No. 12-001-X

# Estimation and inference of domain means subject to qualitative constraints

**Cristian Oliva-Aviles, Mary C. Meyer and Jean D. Opsomer**[1]

## Abstract

In many large-scale surveys, estimates are produced for numerous small domains defined by cross-classifications of demographic, geographic and other variables. Even though the overall sample size of such surveys might be very large, samples sizes for domains are sometimes too small for reliable estimation. We propose an improved estimation approach that is applicable when "natural" or qualitative relationships (such as orderings or other inequality constraints) can be formulated for the domain means at the population level. We stay within a design-based inferential framework but impose constraints representing these relationships on the sample-based estimates. The resulting constrained domain estimator is shown to be design consistent and asymptotically normally distributed as long as the constraints are asymptotically satisfied at the population level. The estimator and its associated variance estimator are readily implemented in practice. The applicability of the method is illustrated on data from the 2015 U.S. National Survey of College Graduates.

**Key Words:** Design-based estimation; Monotone estimation; National Survey of College Graduates.

# 1 Introduction

For many large-scale surveys, a goal is to produce estimates for a large number of domains, many of which might have small sample size. These domains are typically created by cross-classifying categorical variables such as demographic, geographic or other similar characteristics of interest. For instance, the U.S. Current Population Survey releases estimates for domains defined by sex, age, race and/or educational attainment. Similarly, the U.S. American Community Survey produces detailed estimates by sex, age, race/ethnicity for different levels of geography (depending on the release). In another example we will discuss further below, the U.S. National Survey of College Graduates is interested in estimates defined by crossing level and field of degree, occupation and gender. Depending on the survey program, such "granular" estimates are often as important as the higher-level or population estimates.

However, although the overall sample size of such surveys might be very large, samples sizes for numerous domains are often too small for reliable estimates. One possible approach to avoid this problem could be to aggregate small domains into bigger scales so that more reliable direct estimators can be produced for those scales, leading to the generation of more aggregated information than the actual desired scale. An alternative to producing small domain estimates could be changing from a design-based to a model-based estimation methodology such as small area models. While that is certainly a statistically valid approach for creating precise estimates at small scales, it is labor-intensive and sensitive to potential model misspecification. It also replaces the sampling error by model error, so that the mode of inference changes. For those reasons, statistical agencies prefer to stay within the design-based approach, which offers robustness and also allows to stay with the standard mode of inference for surveys.

[1] Cristian Oliva-Aviles, Genentech, Inc.; Mary C. Meyer, Colorado State University; Jean D. Opsomer, Westat, Inc. E-mail: jopsomer@mac.com.

In this paper, we present an estimation approach that is applicable when “natural” or qualitative relationships are expected to hold among the domain means at the population level. These relationships can be used to stabilize the sample domain estimates, while staying within the design-based mode of estimation and inference. The type of relationships we are considering here lead to inequalities among population domain means. For instance, certain job types might be expected to receive better salaries than others, or individuals with graduate degrees in a given discipline are expected to have higher salaries than those without graduate degrees in that discipline. However, given that small domains tend to produce estimates with high variability, such expected population-level relationships are often violated at the sample level. While such violations should be expected by data users due to statistical variability, they might lead them to question the overall reliability of the survey, by producing “absurd” estimates.

There is a large literature in survey statistics related to calibrating survey estimates, see e.g. Särndal, Swensson and Wretman (1992) for an overview. While these estimators also rely on constraints, there are important differences, including the fact that the constraints are equality constraints and that they are applied to the survey weights, not the estimates themselves. While we do not explore this here, it would be possible to combine calibration and constrained estimation, since the latter could use calibrated domain estimates as the starting point for constructing constrained domain estimates. In the model-based setting, Rueda and Lombardía (2012) adapted methods in small area estimation for the case of monotonically ordered domain means.

Recently, Wu, Meyer and Opsomer (2016) proposed a domain mean estimation methodology that relies on the assumption of monotone population domain means along a single domain-defining categorical variable (e.g., age classes). By combining the monotonicity information of domain means and design-based estimators in the estimation stage, they proposed a *constrained* estimator that respects the monotone assumption. Such an estimator was shown to improve precision and variability of domain mean estimates in comparison with direct estimators, given that the assumption of monotonicity is reasonable.

We generalize this work here by allowing a much larger class of constraints between domain means, applicable to the multi-dimensional setting. Many other types of constraints beyond monotonicity may be expected to hold between population domain means in real surveys, especially in the presence of domains defined by the cross-classifications of many categorical variables. In general, any set of linear inequality constraints can be represented through a *constraint matrix*, where each row defines a constraint and each column a domain mean. For illustration of a constraint matrix, suppose the variable of interest is the annual average salary of faculty in land-grant universities of a certain size. Further, consider domains generated from the cross-classification of the variables job position $(x_1;\ 1 = \text{Untenured and } 2 = \text{Tenured})$ and three specific departments $(x_2;\ 1 = \text{Anthropology},\ 2 = \text{English and } 3 = \text{Engineering})$. Under the assumptions that, on average within a discipline, tenured faculty have higher salaries than untenured faculty; and that, within tenured and untenured, Engineering faculty members are expected to have higher salaries than those in either the Anthropology or English departments, then we can express the corresponding restrictions as,

$$\mathbf{A}\boldsymbol{\mu} \geq \mathbf{0}, \quad \text{where} \quad \mathbf{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}, \tag{1.1}$$

$\boldsymbol{\mu} = (\mu_{11}, \mu_{21}, \mu_{12}, \mu_{22}, \mu_{13}, \mu_{23})^{\top}$, with $\mu_{ij}$ representing the mean of the domain that corresponds to $x_1 = i$ and $x_2 = j$; $\mathbf{0}$ being the zero vector, and the inequality being element-wise. This paper describes a new constrained estimator for population domain means that respect constraints that can be expressed with matrix inequalities of the form given in (1.1). By combining design-based domain mean estimators with these shape constraints, we propose a broadly applicable estimator that improves precision and variability of the most common direct estimators.

The remainder of the paper is organized as follows. In Section 2 we formally introduce the constrained estimator and propose a linearization-based method for variance estimation. This section also contains some scenarios of interest where shape constraints can naturally arise for survey data. Section 3 states the main theoretical properties of the constrained estimator. The necessary assumptions used in these theoretical derivations are also stated in this section. Proofs of main theorems and auxiliary lemmas are provided in the Appendix. Section 4 shows through simulations that the constrained estimator improves domain mean estimation and variability in comparison with the unconstrained estimator, even when the assumed shape holds only approximately at the population level. Section 5 demonstrates the advantages of the proposed methodology on real survey data through an application to the 2015 National Survey of College Graduates. A few concluding remarks are provided in Section 6.

# 2 Constrained estimation and inference for domain means

## 2.1 Notation and preliminaries

Let $U_N$ be the set of elements in a population of size $N$. Consider a sample $s_N$ of size $n_N$ that is drawn from $U_N$ using a probability sampling design $p_N(\cdot)$. Denote $\pi_{k,N} = \Pr(k \in s_N)$ and $\pi_{kl,N} = \Pr(k \in s_N, l \in s_N)$ as the first and second order inclusion probabilities, respectively. Assume that $\pi_{k,N} > 0$, $\pi_{kl,N} > 0$ for $k, l \in U_N$. To simplify notation, we will adopt the usual convention of suppressing the subscript $N$ unless it is needed for clarity. Denote $\{U_d\}_{d=1}^{D}$ as a domain partition of $U$, where $D$ is the number of domains and each $U_d$ is of size $N_d$. Also, let $s_d$ be the subset of size $n_d$ of $s$ that belongs to $U_d$.

For any study variable $y$, $\bar{\mathbf{y}}_U = \left(\bar{y}_{U_1}, \ldots, \bar{y}_{U_D}\right)^{\top}$ denotes the vector of population domain means, where

$$\bar{y}_{U_D} = \frac{\sum_{k \in U_d} y_k}{N_d}. \tag{2.1}$$

We will focus on the Hájek estimator of $\bar{y}_{U_D}$, given by

$$\tilde{y}_{s_d} = \frac{\sum_{k \in s_d} y_k / \pi_k}{\hat{N}_d} \tag{2.2}$$

with $\hat{N}_d = \sum_{k \in s_d} 1/\pi_k$, and let $\tilde{\mathbf{y}}_s$ to be the vector of estimators. The results will also hold for the Horvitz-Thompson estimator with minor modifications, but it will not be explicitly addressed in what follows.

## 2.2 Proposed estimator

Assume there is information available regarding relationships between the population domain means that can be expressed with $m$ constraints through a $m \times D$ *irreducible* constraint matrix $\mathbf{A}$. A matrix $\mathbf{A}$ is irreducible if none of its rows is a positive linear combination of other rows, and if the origin is also not a positive linear combination of its rows (Meyer, 1999). In practical terms, this means that there are no redundant constraints in $\mathbf{A}$. To take advantage of $\tilde{\mathbf{y}}_s$ to obtain an estimator that respects these shape constraints, we propose the constrained estimator $\tilde{\boldsymbol{\theta}}_s = \left(\tilde{\theta}_{s_1}, \ldots, \tilde{\theta}_{s_D}\right)^{\top}$ to be the unique vector that solves the following constrained weighted least squares problem,

$$\min_{\boldsymbol{\theta}} (\tilde{\mathbf{y}}_s - \boldsymbol{\theta})^{\top} \mathbf{W}_s (\tilde{\mathbf{y}}_s - \boldsymbol{\theta}) \quad \text{subject to} \quad \mathbf{A}\boldsymbol{\theta} \geq \mathbf{0}; \tag{2.3}$$

where $\mathbf{W}_s$ is the diagonal matrix with elements $\hat{N}_1/\hat{N}$, $\hat{N}_2/\hat{N}$, $\ldots$, $\hat{N}_D/\hat{N}$, and $\hat{N} = \sum_{d=1}^{D} \hat{N}_d$. The constrained problem in equation (2.3) can be alternatively written as finding the unique vector $\tilde{\boldsymbol{\phi}}_s$ that solves

$$\min_{\boldsymbol{\phi}} \| \tilde{\mathbf{z}}_s - \boldsymbol{\phi} \|^2 \quad \text{subject to} \quad \mathbf{A}_s \boldsymbol{\phi} \geq \mathbf{0}, \tag{2.4}$$

where $\tilde{\mathbf{z}}_s = \mathbf{W}_s^{1/2} \tilde{\mathbf{y}}_s$, $\boldsymbol{\phi} = \mathbf{W}_s^{1/2} \boldsymbol{\theta}$, and $\mathbf{A}_s = \mathbf{A}\mathbf{W}_s^{-1/2}$. The transformed constrained matrix $\mathbf{A}_s$ is also irreducible if $\mathbf{A}$ is, and it depends on the sample although $\mathbf{A}$ does not. The solution $\tilde{\boldsymbol{\phi}}_s$ is the *projection* of $\tilde{\mathbf{z}}_s$ onto the set of vectors $\boldsymbol{\phi}$ that satisfy the condition $\mathbf{A}_s \boldsymbol{\phi} \geq \mathbf{0}$. This set is a polyhedral convex cone, called the *constraint cone* $\Omega_s$ defined by $\mathbf{A}_s$; specifically,

$$\Omega_s = \left\{ \boldsymbol{\phi} \in \mathbb{R}^D \text{: } \mathbf{A}_s \boldsymbol{\phi} \geq \mathbf{0} \right\}. \tag{2.5}$$

We use the notation $\tilde{\boldsymbol{\phi}}_s = \Pi(\tilde{\mathbf{z}}_s \mid \Omega_s)$, where $\Pi(\mathbf{u} \mid \mathcal{S})$ stands for the projection of $\mathbf{u}$ onto the set $\mathcal{S}$, i.e., the closest vector in $\mathcal{S}$ to $\mathbf{u}$.

Projections onto such cones are well understood; see Rockafellar (1970) or Meyer (1999) for details. In terms of this work, the main results from cone projection theory are summarized here. The cone can be characterized by a set of *edges* generating the cone; that is, a vector is in the cone if and only if it is a linear combination of the edges with non-negative coefficients. (Picture a pyramid with vertex at the origin, extending out indefinitely.) Subsets of the edges define the *faces* of the cone, and the projection of $\tilde{\mathbf{z}}_s$ onto the cone lands on one of the faces. Once the edges defining this face are determined, the projection can be characterized as an ordinary least-squares projection onto the linear space spanned by this subset of edges. This property is crucial for both the algorithm for projection and for inference, because the projection onto the cone can be characterized as a linear projection.

For this work, we will project $\tilde{\mathbf{z}}_s$ onto the *polar cone* $\Omega_s^0$ (Rockafellar, 1970, page 121), defined as

$$\Omega_s^0 = \{\boldsymbol{\rho} \in \mathbb{R}^D \colon \langle \boldsymbol{\rho}, \boldsymbol{\phi} \rangle \le 0, \ \forall \boldsymbol{\phi} \in \Omega_s\}, \tag{2.6}$$

where $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$. That is, the polar cone is the set of vectors that form obtuse angles with all vectors in $\Omega_s$. The polar cone is analogous to the orthogonal space in linear least-squares projections, in that the projection of a vector onto the polar cone is the residual of its projection onto the constraint cone, and vice-versa. Meyer (1999) showed that the negative rows of an irreducible matrix are the *edges* (generators) of the polar cone, leading to the following characterization of the polar cone in (2.6):

$$\Omega_s^0 = \left\{\boldsymbol{\rho} \in \mathbb{R}^D \colon \boldsymbol{\rho} = \sum_{j=1}^{m} a_j \boldsymbol{\gamma}_{s_j}, \ a_j \ge 0, \ j = 1, 2, \ldots, m\right\}, \tag{2.7}$$

where $\boldsymbol{\gamma}_{s_1}, \boldsymbol{\gamma}_{s_2}, \ldots, \boldsymbol{\gamma}_{s_m}$ are the rows of $-\mathbf{A}_s$. Robertson, Wright and Dykstra (1988, page 17) established necessary and sufficient conditions for a vector $\tilde{\boldsymbol{\phi}}_s$ to be the projection of $\tilde{\mathbf{z}}_s$ onto $\Omega_s$. That is, $\tilde{\boldsymbol{\phi}}_s \in \Omega_s$ solves the constrained problem in (2.4) if and only if

$$\langle \tilde{\mathbf{z}}_s - \tilde{\boldsymbol{\phi}}_s, \tilde{\boldsymbol{\phi}}_s \rangle = 0, \quad \text{and} \quad \langle \tilde{\mathbf{z}}_s - \tilde{\boldsymbol{\phi}}_s, \boldsymbol{\phi} \rangle \le 0, \ \forall \boldsymbol{\phi} \in \Omega_s.$$

Moreover, the above conditions can be adapted to the polar cone as follows: the vector $\tilde{\boldsymbol{\rho}}_s \in \Omega_s^0$ minimizes $\| \tilde{\mathbf{z}}_s - \boldsymbol{\rho} \|^2$ over $\Omega_s^0$ if and only if

$$\langle \tilde{\mathbf{z}}_s - \tilde{\boldsymbol{\rho}}_s, \tilde{\boldsymbol{\rho}}_s \rangle = 0, \quad \text{and} \quad \langle \tilde{\mathbf{z}}_s - \tilde{\boldsymbol{\rho}}_s, \boldsymbol{\gamma}_{s_j} \rangle \le 0 \quad \text{for} \quad j = 1, 2, \ldots, m. \tag{2.8}$$

The conditions in (2.8) can be used to show that the projection of $\tilde{\mathbf{z}}_s$ onto the polar cone $\Omega_s^0$ coincides with the projection onto the linear space generated by the edges $\boldsymbol{\gamma}_{s_j}$ such that $\langle \tilde{\mathbf{z}}_s - \tilde{\boldsymbol{\rho}}_s, \boldsymbol{\gamma}_{s_j} \rangle = 0$. This set of edges could be empty, meaning that the projection onto $\Omega_s^0$ is equal to the projection onto the zero vector. In that case, the unconstrained minimum satisfies all the constraints. Alternatively, this set of edges might not be unique. To formalize these ideas, denote $V_{s,J} = \{\boldsymbol{\gamma}_{s_j} \colon j \in J\}$ for any $J \subseteq \{1, 2, \ldots, m\}$. Define the set $\mathcal{F}_{s,J}$ as,

$$\bar{\mathcal{F}}_{s,J} = \left\{ \boldsymbol{\rho} \in \mathbb{R}^D \colon \boldsymbol{\rho} = \sum_{j \in J} a_j \boldsymbol{\gamma}_{s_j},\ \ a_j \geq 0,\ \ j \in J \right\}, \tag{2.9}$$

where $\bar{\mathcal{F}}_{s,\varnothing} = \mathbf{0}$ by convention. (Technically, this set is the closure of a face of the cone.) That is, $\bar{\mathcal{F}}_{s,J}$ is a closed polyhedral sub-cone of $\Omega_s^0$ that starts at the origin and is defined by the edges in $V_{s,J}$. Further, let $\mathcal{L}(V_{s,J})$ be the linear space generated by the vectors in $V_{s,J}$. It is shown in Meyer (1999) that projecting onto $\Omega_s^0$ is equivalent to projecting onto $\mathcal{L}(V_{s,J})$, for an appropriate set $J$. If the rows of the constraint matrix $\mathbf{A}$ are linearly independent, then the minimal set $J$ is unique; otherwise there may be more than one $J$ that defines the linear space. In the latter case, however, the projection is still unique (see Theorem 1 of the next section).

Wu et al. (2016) considered the solution to (2.3), in the special case of a monotone relationship between domains defined along a single categorical variable. In that case, the solution is equivalent to that of the Pooled Adjacent Violator Algorithm (PAVA), which has an explicit expression in terms of a pooling of neighboring domains. The theoretical results in Wu et al. (2016) were obtained using that explicit expression, and hence do not apply to the more general setting considered here. Nevertheless, as was the case with the simple 6-domain example in Section 1 and in many situations of practical interest, the specific matrix $\mathbf{A}$ will often correspond to a multivariate *partial ordering* of the domain means. Under partial ordering, the solution to the constrained minimization in (2.3) is again equivalent to a pooling of neighboring domains in such a way that the partial order constraints are respected. See for instance Robertson et al. (1988, page 23) for an explicit expression of this pooled domain expression under partial ordering, including the definition of the pooling. However, unlike PAVA in the univariate case, this does not lead to a practical general computational algorithm. In the current paper, we will allow for arbitrary irreducible constraint matrix $\mathbf{A}$, which will include partial ordering and univariate monotonicity as special cases.

One possible general approach to computing $\tilde{\boldsymbol{\phi}}_s$ is based on the edges of the constraint cone $\Omega_s$. However, the number of edges can be considerably larger than the number of constraints for large values of $D$, especially for the case when there are more constraints than domains (see Meyer, 1999). Moreover, given the lack of a general closed form solution for the edges of $\Omega_s$ (when $m > D$), the edges need to be computed numerically in that case. This task is computationally demanding, which makes this approach an inefficient way to compute $\tilde{\boldsymbol{\phi}}_s$. A more efficient algorithm based on computing the projection onto the polar cone has been developed: the Cone Projection Algorithm (CPA) (Meyer, 2013). This alternative approach takes advantage of the easy-to-find edges $\boldsymbol{\gamma}_{s_j}$ of the polar cone, the conditions in (2.8), and the fact that $\Pi(\tilde{\mathbf{z}}_s \,|\, \Omega_s) = \tilde{\mathbf{z}}_s - \Pi(\tilde{\mathbf{z}}_s \,|\, \Omega_s^0)$. The latter fact is a key component on the proofs of the main theoretical results shown in this paper. CPA has been implemented in the software R into the `coneproj` package. See Liao and Meyer (2014) for further details.

For the situations in which the constraints correspond to complete or partial ordering, the CPA solution once again corresponds to domain pooling. After this, the domain mean estimates can be explicitly computed as sample-based domain means for the CPA-determined pooled domains. This greatly facilitates incorporating this methodology into survey estimation practice, because the pooled domain definitions can be readily communicated as part of the instructions accompanying a survey dataset release, and the estimates can be calculated without requiring access to specialized software.

## 2.3 Variance estimation of $\tilde{\theta}_{s_d}$

Estimating appropriately the variance of $\tilde{\theta}_{s_d}$ is a complicated task, derived from the fact that the projection of $\tilde{\mathbf{z}}_s$ onto $\Omega_s^0$ (or onto $\Omega_s$) might not always land on the same linear space $\mathcal{L}(V_{s,J})$ for different samples $s$. To better understand that, we define $\mathcal{G}_s$ as the set of all subsets $J \subseteq \{1, 2, \ldots, m\}$ such that $\Pi(\tilde{\mathbf{z}}_s \mid \Omega_s^0) = \Pi(\tilde{\mathbf{z}}_s \mid \mathcal{L}(V_{s,J})) \in \bar{\mathcal{F}}_{s,J}$, as defined in (2.9). As noted earlier, there could be different sets $J_1$ and $J_2$ such that the projection onto the polar cone $\Omega_s^0$ is equal to projecting onto either $\mathcal{L}(V_{s,J_1})$ or $\mathcal{L}(V_{s,J_2})$. However, independently of which set is chosen, the projection $\tilde{\boldsymbol{\rho}}_s$ is unique.

To illustrate the above point, consider the following restrictions when there are only 3 domains: the first domain mean is expected to be at the most equal to the second domain mean, and the third domain mean is expected to be at least equal to the average of the first two domain means. Hence, the constraint matrix $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & -1 & 2 \end{pmatrix}.$$

Suppose it is observed that $\tilde{y}_{s_1} = \tilde{y}_{s_2} < \tilde{y}_{s_3}$. The transformed vector $\tilde{\mathbf{z}}_s$ has elements of the form

$$\tilde{z}_{s_1} = \sqrt{\frac{\hat{N}_1}{\hat{N}}}\,\tilde{y}_{s_1}, \quad \tilde{z}_{s_2} = \sqrt{\frac{\hat{N}_2}{\hat{N}}}\,\tilde{y}_{s_2}, \quad \tilde{z}_{s_3} = \sqrt{\frac{\hat{N}_3}{\hat{N}}}\,\tilde{y}_{s_3}.$$

In this setting, it is straightforward to see that $\Pi(\tilde{\mathbf{z}}_s \mid \Omega_s^0) = \mathbf{0}$. In the process of computing it using the general algorithm, we project $\tilde{\mathbf{z}}_s$ onto each of the $2^2 = 4$ linear spaces generated by the polar cone edges

$$\boldsymbol{\gamma}_{s_1} = \left(\sqrt{\frac{\hat{N}}{\hat{N}_1}}, -\sqrt{\frac{\hat{N}}{\hat{N}_2}}, 0\right)^{\top}, \quad \boldsymbol{\gamma}_{s_2} = \left(\sqrt{\frac{\hat{N}}{\hat{N}_1}}, \sqrt{\frac{\hat{N}}{\hat{N}_2}}, -2\sqrt{\frac{\hat{N}}{\hat{N}_3}}\right)^{\top}.$$

Hence, it can be seen that the conditions $\Pi(\tilde{\mathbf{z}}_s \mid \Omega_s^0) = \mathbf{0} = \Pi(\tilde{\mathbf{z}}_s \mid \mathcal{L}(V_{s,J})) \in \bar{\mathcal{F}}_{s,J}$ are satisfied only for $J = \varnothing$ and $J = \{1\}$, which implies that $\mathcal{G}_s = \{\varnothing, \{1\}\}$. Moreover, note that $V_{s,\varnothing}$ and $V_{s,\{1\}}$ do not span the same linear spaces, which is what complicates the variance estimation of $\tilde{\theta}_{s_d}$. In the model-based case with continuous variables, the set of sample vectors where these scenarios occur has measure zero. However, they cannot be excluded in the design-based setting.

We propose a variance estimator for $\tilde{\theta}_{s_d}$ that relies on the sets in $\mathcal{G}_s$ and is based on linearization methods. Consider any fixed set $J \in \mathcal{G}_s$, and let $\mathbf{P}_{s,J}$ be the projection matrix corresponding to the linear space $\mathcal{L}(V_{s,J})$, where $\mathbf{P}_{s,\varnothing}$ is the matrix of zeros by convention. By the selection of $J$, then $\tilde{\boldsymbol{\rho}}_s$

can be expressed as $\mathbf{P}_{s,J}\tilde{\mathbf{z}}_s$, which implies that $\tilde{\boldsymbol{\theta}}_s$ can be written as $\tilde{\boldsymbol{\theta}}_{s,J} = \tilde{\mathbf{y}}_s - \mathbf{W}_s^{-1/2}\mathbf{P}_{s,J}\mathbf{W}_s^{1/2}\tilde{\mathbf{y}}_s$, where we add the subscript $J$ in $\tilde{\boldsymbol{\theta}}_s$ to be aware that the expression depends on the chosen $J$.

Now, observe that $\tilde{\boldsymbol{\theta}}_{s,J}$ is a smooth non-linear function of the $\hat{t}_d$'s and the $\hat{N}_d$'s, where $\hat{t}_d$ is the Horvitz-Thompson estimator of $t_d = \sum_{k \in U_d} y_k$. Therefore, treating $J$ as fixed, we obtain the asymptotic variance of $\tilde{\theta}_{s_d,J}$ via Taylor linearization (Särndal et al., 1992, page 175) as

$$\mathrm{AV}\left(\tilde{\theta}_{s_d,J}\right) = \sum_{k \in U}\sum_{l \in U} \Delta_{kl}\frac{u_k}{\pi_k}\frac{u_l}{\pi_l}, \tag{2.10}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$, and

$$u_k = \sum_{i=1}^{D}\alpha_i\, y_k 1_{k \in U_i} + \sum_{i=1}^{D}\beta_i 1_{k \in U_i} \quad \text{for} \quad k = 1,\ 2,\ \ldots,\ N,$$

with $1_A$ being the indicator variable for the event $A$, and

$$\alpha_i = \left.\frac{\partial\tilde{\theta}_{s_d,J}}{\partial\hat{t}_i}\right|_{(\hat{t}_1,\ldots,\hat{t}_D,\hat{N}_1,\ldots,\hat{N}_D) = (t_1,\ldots,t_D,N_1,\ldots,N_D)}; \quad \beta_i = \left.\frac{\partial\tilde{\theta}_{s_d,J}}{\partial\hat{N}_i}\right|_{(\hat{t}_1,\ldots,\hat{t}_D,\hat{N}_1,\ldots,\hat{N}_D) = (t_1,\ldots,t_D,N_1,\ldots,N_D)}.$$

In addition, a consistent estimator of the asymptotic variance in (2.10) is given by

$$\hat{V}\left(\tilde{\theta}_{s_d,J}\right) = \sum_{k \in s}\sum_{l \in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{\hat{u}_k}{\pi_k}\frac{\hat{u}_l}{\pi_l}, \tag{2.11}$$

where

$$\hat{u}_k = \sum_{i=1}^{D}\hat{\alpha}_i\, y_k 1_{k \in s_i} + \sum_{i=1}^{D}\hat{\beta}_i 1_{k \in s_i} \quad \text{for} \quad k = 1,\ 2,\ \ldots,\ N,$$

with $\hat{\alpha}_i$, $\hat{\beta}_i$ obtained from $\alpha_i$, $\beta_i$ by substituting the appropriate Horvitz-Thompson estimators for each population total. We propose the estimator in (2.11), computed at the $J$ obtained in the sample, as a variance estimator of $\tilde{\theta}_{s_d}$.

To provide a clear example of the proposed variance estimator for $\tilde{\theta}_{s_d}$, consider the setting presented at the beginning of this subsection. Since $\mathcal{G}_s = \{\varnothing,\ \{1\}\}$, it might be of interest to compute the estimated variance of $\tilde{\theta}_{s_d,J}$ for $J = \{1\}$ and certain $d$. The matrix $\mathbf{P}_{s,\{1\}}$ is the projection matrix corresponding to the linear space generated by $\boldsymbol{\gamma}_{s_1}$, given by

$$\mathbf{P}_{s,\{1\}} = \left(\hat{N}_1 + \hat{N}_2\right)^{-1}\begin{pmatrix} \hat{N}_2 & -\sqrt{\hat{N}_1\hat{N}_2} & 0 \\ -\sqrt{\hat{N}_1\hat{N}_2} & \hat{N}_1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Note that $\mathbf{P}_{s,\{1\}}$ is a function of $\left(\hat{N}_1,\ \hat{N}_2,\ \hat{N}_3\right)$ because $\boldsymbol{\gamma}_{s_1}$ is. Using the above equation, $\tilde{\boldsymbol{\theta}}_{s,\{1\}}$ can be simplified to the following expression,

$$\tilde{\boldsymbol{\theta}}_{s,\{1\}} = \left(\tilde{\theta}_{s_1,\{1\}},\ \tilde{\theta}_{s_2,\{1\}},\ \tilde{\theta}_{s_3,\{1\}}\right)^{\top} = \left(\frac{\hat{N}_1\,\tilde{y}_{s_1} + \hat{N}_2\,\tilde{y}_{s_2}}{\hat{N}_1 + \hat{N}_2},\ \frac{\hat{N}_1\,\tilde{y}_{s_1} + \hat{N}_2\,\tilde{y}_{s_2}}{\hat{N}_1 + \hat{N}_2},\ \tilde{y}_{s_3}\right)^{\top}$$

$$= \left(\frac{\hat{t}_1 + \hat{t}_2}{\hat{N}_1 + \hat{N}_2},\ \frac{\hat{t}_1 + \hat{t}_2}{\hat{N}_1 + \hat{N}_2},\ \frac{\hat{t}_3}{\hat{N}_3}\right)^{\top}.$$

Therefore, given a domain $d,$ the $\alpha$'s and $\beta$'s can be derived by taking the partial derivatives of $\tilde{\theta}_{s_d,\{1\}}$ with respect to the $\hat{t}$'s and $\hat{N}$'s, and evaluating such derivatives at the $t$'s and $N$'s. For $d = 2,$ that is,

$$\alpha_1 = \alpha_2 = \frac{1}{N_1 + N_2}, \qquad \alpha_3 = 0,$$

$$\beta_1 = \beta_2 = -\frac{t_1 + t_2}{\left(N_1 + N_2\right)^2},\ \beta_3 = 0.$$

The $\hat{\alpha}$'s and $\hat{\beta}$'s are computed by substituting Horvitz-Thompson estimators in the above equations, which are then used to evaluate $\hat{u}_k$ for each $k$ in the sample $s.$ Finally, the proposed variance estimator in (2.11) can be computed.

# 3 Properties of the constrained estimator

## 3.1 Assumptions

To derive our theoretical results, we make assumptions on the asymptotic behavior of the population $U_N$ and the sampling design $p_N$:

A1. The number of domains $D$ is fixed.

A2. $\limsup_{N\to\infty} N^{-1}\sum_{k\in U}|y_k|^r < \infty,$ for $r = 1,\ 2.$

A3. For $d = 1, \ldots, D,$ there exist constants $\mu_d$ and $r_d > 0$ such that $\bar{y}_{U_d,N} - \mu_d = O\left(N^{-1/2}\right)$ and $N_{d,N}/N - r_d = O\left(N^{-1/2}\right),$ for all $d.$

A4. The sample size $n_N$ is non-random and satisfies $0 < \lim_{N\to\infty} n_N/N < 1.$ In addition, there exists $\varepsilon,\ 0 < \varepsilon < 1,$ such that $n_{d,N} \geq \varepsilon n_N/D$ for all $d$ and all $N.$

A5. For all $N,$ $\min_{k\in U_N}\pi_k \geq \lambda > 0,$ $\min_{k,l\in U_N}\pi_{kl} \geq \lambda^* > 0,$ and

$$\limsup_{N\to\infty} n_N \max_{k,l\in U_N:\ k\neq l}|\Delta_{kl}| < \infty.$$

A6. The Horvitz-Thompson estimator $\hat{\mathbf{x}}_{s_N}$ of the $2D$-dimensional vector of population means $\overline{\mathbf{x}}_{U_N} = N^{-1}(t_1, \ldots, t_D, N_1, \ldots, N_D)^{\top}$ satisfies

$$\operatorname{var}_{p_N}\left(\hat{\mathbf{x}}_{s_N}\right)^{-1/2}\left(\hat{\mathbf{x}}_{s_N} - \overline{\mathbf{x}}_{U_N}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{I}_{2D}\right),$$

and

$$\widehat{\operatorname{var}}\left(\hat{\mathbf{x}}_{s_N}\right) - \operatorname{var}_{p_N}\left(\hat{\mathbf{x}}_{s_N}\right) = o_p\left(n_N^{-1}\right);$$

where $\mathbf{I}_q$ denotes the identity matrix of dimension $q$, the design variance-covariance matrix $\operatorname{var}_{p_N}\left(\hat{\mathbf{x}}_{s_N}\right)$ is positive definite, and $\widehat{\operatorname{var}}\left(\hat{\mathbf{x}}_{s_N}\right)$ is the Horvitz-Thompson estimator of $\operatorname{var}_{p_N}$.

Assumption A1 establishes that the number of domains remains constant as the population size changes. The condition in Assumption A2 is made to ensure design consistency of Horvitz-Thompson estimators at the population and domain levels. In particular, note that this condition is satisfied when the variable $y$ is bounded, which can be naturally assumed for many types of survey variables. Assumption A3 guarantees that the population domain means and sizes converge to the limiting values $\mu_d$ and $r_d$, respectively. Alternatively, the $\mu$ values can be thought as superpopulation expectations for a distribution that generates the population elements $y_k$ as independent draws. In fact, our theoretical results depend on whether the assumed constraints hold for these superpopulation expectations and not for the population domain means. Although this might seem to be inappropriate given our interest on using constraints at the population level, Assumption A3 ensures that the shape of the domain means would be reasonably close to the shape of the superpopulation means. Assumption A4 states that the sample size in each domain cannot be smaller than a fraction of the ratio $n/D$, which would be obtained by dividing equally the sample size over all domains. This assumption aims to ensure that the moments of smooth functions of the $N^{-1}\hat{t}_d$ and the $N^{-1}\hat{N}_d$ are bounded. Also, it assumes that the sample size is non-random. This can be adapted to a random sample size by imposing certain conditions on the expected sample size $\mathrm{E}_p(n)$. Assumption A5 establishes non-zero lower bounds for both first and second order inclusion probabilities, and states that the design covariances $\Delta_{kl}$ must converge to zero at least as fast as $n^{-1}$. Assumption A6 ensures asymptotic normality for $\hat{\mathbf{x}}_{s_N}$, which is needed to maintain normality properties on non-linear estimators that are expressed as smooth functions of $\hat{\mathbf{x}}_{s_N}$. It is also used to establish consistency conditions on the variance-covariance estimator. For specific designs, asymptotic normality results are available in the literature, including the classical result by Hájek (1960) for Poisson sampling and simple random sampling without replacement. Additional central limit theorems for stratified sampling include Krewski and Rao (1981), who considered stratified unequal probability samples with replacement, Bickel and Freedman (1984), who considered stratified simple random sampling without replacement, and Breidt, Opsomer and Sanchez-Borrego (2016), who considered general unequal probability designs, with or without replacement.

## 3.2 Main results

We derive the theoretical properties of the constrained estimator by focusing on the projection onto $\Omega_s^0$ instead of $\Omega_s$. Recall that the edges of the polar cone $\Omega_s^0$ are simply the $m$ rows of $-\mathbf{A}_s$, denoted by $\boldsymbol{\gamma}_{s_j}$; and that $\tilde{\boldsymbol{\rho}}_s$, the projection onto $\Omega_s^0$, can be described by the sets $J \in \mathcal{G}_s$. Being able to characterize the property that $J \in \mathcal{G}_s$ in terms of the vectors in $V_{s,J}$ allow us to obtain theoretical convergence rates, which are used to develop inference properties of the constrained estimator. When the set $J \in \mathcal{G}_s$ produces a set of linear independent vectors $V_{s,J}$, then it is straightforward that $\tilde{\boldsymbol{\rho}}_s$ can be written as $\mathbf{P}_{s,J}\,\tilde{\mathbf{z}}_s = \mathbf{A}_{s,J}^{\top}\left(\mathbf{A}_{s,J}\mathbf{A}_{s,J}^{\top}\right)^{-1}\mathbf{A}_{s,J}\,\tilde{\mathbf{z}}_s$, where $\mathbf{A}_{s,J}$ denotes the matrix formed by the rows of $\mathbf{A}_s$ in positions $J$. Hence, based on the conditions in (2.8), $J \in \mathcal{G}_s$ if and only if

$$\left\langle \tilde{\mathbf{z}}_s - \mathbf{P}_{s,J}\,\tilde{\mathbf{z}}_s,\ \boldsymbol{\gamma}_{s_j}\right\rangle \le 0 \quad \text{for} \quad j \notin J, \quad \text{and} \quad \left(\mathbf{A}_{s,J}\,\mathbf{A}_{s,J}^{\top}\right)^{-1}\mathbf{A}_{s,J}\,\tilde{\mathbf{z}}_s \ge \mathbf{0} \tag{3.1}$$

in this case, where the latter condition assures that $\Pi\left(\tilde{\mathbf{z}}_s \,|\, \mathcal{L}\left(V_{s,J}\right)\right) \in \bar{\mathcal{F}}_{s,J}$. However, it is possible that the set $J \in \mathcal{G}_s$ produces a set of linearly dependent vectors $V_{s,J}$. In that case, Theorem 1 below guarantees that it is always possible to find a subset $J^* \subset J$ such that $V_{s,J^*}$ is a linearly independent set that spans the same linear space as $V_{s,J}$ and that satisfies $J^* \in \mathcal{G}_s$. Thus, analogous conditions as in (3.1) can be established using $J^*$ instead of $J$.

**Theorem 1.** *Let $\mathbf{A}$ be a $m \times D$ irreducible matrix with rows $-\boldsymbol{\gamma}_j$. Let $\Omega^0$ be its corresponding polar cone. For any set $J \subseteq \{1, 2, \ldots, m\}$, define $V_J = \{\boldsymbol{\gamma}_j\colon j \in J\}$. Further, denote $\bar{\mathcal{F}}_J$ to be the subcone of $\Omega^0$ generated by the edges given by the set $J$. For a vector $\mathbf{z}$, define its set $\mathcal{G}$ to be formed by all sets $J \subseteq \{1, 2, \ldots, m\}$ such that $\Pi(\mathbf{z}|\Omega^0) = \Pi(\mathbf{z}|\mathcal{L}(V_J)) \in \bar{\mathcal{F}}_J$. Suppose $J$ is a non-empty set such that $V_J$ is a linearly dependent set and $J \in \mathcal{G}$. Then, there exists $J^* \subset J$ such that $V_{J^*}$ is a linearly independent set, $\mathcal{L}\left(V_{J^*}\right) = \mathcal{L}\left(V_J\right)$, and $J^* \in \mathcal{G}$.*

All above concepts that have been defined at the sample level can be analogously defined at the superpopulation level. In particular, let $\mathcal{G}_\mu$ be the set of all subsets $J \subseteq \{1, \ldots, m\}$ such that $\Pi\left(\mathbf{z}_\mu \,|\, \Omega_\mu^0\right) = \Pi\left(\mathbf{z}_\mu \,|\, \mathcal{L}\left(V_{\mu,J}\right)\right) \in \bar{\mathcal{F}}_{\mu,J}$, where $\mathbf{z}_\mu, \Omega_\mu^0, V_{\mu,J}$ and $\bar{\mathcal{F}}_{\mu,J}$ are the analogous versions of $\tilde{\mathbf{z}}_s, \Omega_s^0, V_{s,J}$ and $\bar{\mathcal{F}}_{s,J}$ obtained by substituting $\tilde{\mathbf{y}}_s$ and $\mathbf{W}_s$ by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)$ and $\mathbf{W}_\mu = \operatorname{diag}(r_1, r_2, \ldots, r_D)$. Necessary and sufficient conditions as in (2.8) can be analogously established to characterize the vector $\boldsymbol{\rho}_\mu$ to be the projection onto $\Omega_\mu^0$.

Recall the set $\mathcal{G}_s$ could vary for different samples. Also, note that highly variable small samples are likely to choose sets $J \in \mathcal{G}_s$ that are not chosen in the “asymptotically correct” $\mathcal{G}_\mu$. However, as the sample size increases, these incorrect choices are less likely to occur since the sample domain means get closer to the limiting population domain means. This idea is made more precise in Theorem 2, which states that sets that are not in $\mathcal{G}_\mu$ have an asymptotically negligible probability of being chosen in the sample.

**Theorem 2.** *Consider any set $J \subseteq \{1, 2, \ldots, m\}$ such that $J \notin \mathcal{G}_\mu$. Then, $P(J \in \mathcal{G}_s) = O(n^{-1})$.*

Theorem 3 below shows the asymptotic normality of the constrained estimator and justifies the use of the linearization-based variance estimator for the observed projection (or pooling, in the case of partial ordering) for asymptotic inference for the finite population domain mean. This generalizes Theorem 2 of Wu et al. (2016), where only monotone restrictions were considered. Note the presence of a bias term $B$ in the mean of the asymptotic distribution. This undesirable situation occurs when there is more than one set $J \in \mathcal{G}_{\mu}$ such that their corresponding edges in $V_{\mu, J}$ span different linear spaces, or equivalently, that the projection onto the polar cone $\Omega_{\mu}^{0}$ belongs to the intersection of those different linear spaces. However, when the constraints hold strictly, i.e., $\mathbf{A\mu} > \mathbf{0}$, the vector $\mathbf{z}_{\mu}$ is strictly inside the constraint cone $\Omega_{\mu}$, and in this case there is no set $J \neq \varnothing$ such that $\Pi(\mathbf{z}_{\mu} | \mathcal{L}(V_{\mu, J})) = \mathbf{0}$. Thus, in this case, the bias term vanishes.

**Theorem 3.** *Suppose that* $\boldsymbol{\mu}$ *satisfies* $\mathbf{A\mu} \geq \mathbf{0}$. *Consider any set* $J$ *such that* $J \in \mathcal{G}_{s}$. *Then*

$$\hat{V}\left(\tilde{\theta}_{s_d, J}\right)^{-1/2}\left(\tilde{\theta}_{s_d} - \bar{y}_{U_d}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(B, 1),$$

*for any* $d = 1, 2, \ldots, D$, *where* $B = O\left(\sqrt{\frac{n}{N}}\right)$ *is a bias term that vanishes when* $\mathbf{A\mu} > \mathbf{0}$.

Theorem 3 relies on the fact that the assumed shape constraints hold for the vector of limiting domain means $\boldsymbol{\mu}$ instead of for the vector of population domain means $\bar{\mathbf{y}}_{U}$. In the next section, we show through simulations that the constrained estimator improves both estimation and variability when the population domains are approximately close to the assumed shape, in comparison with unconstrained estimators.

# 4 Performance of constrained estimator

## 4.1 Simulations

We run simulation experiments to measure the performance of the proposed methodology to carry out estimation and inference of population domain means. Given a pair of natural numbers $D_1$ and $D_2$, we generate the limiting domain means $\mu_d$ from the monotone bivariate function $\mu(x_1, x_2)$ given by

$$\mu(x_1, x_2) = \sqrt{1 + 4x_1/D_1} + \frac{4\exp(0.5 + 2x_2/D_2)}{1 + \exp(0.5 + 2x_2/D_2)}.$$

The $\mu_d$ are created by evaluating $\mu(x_1, x_2)$ at every combination of $x_1 = 1, 2, \ldots, D_1$ and $x_2 = 1, 2, \ldots, D_2$, producing a total number of domains equal to $D = D_1 D_2$. We set $D_1 = 6$ and $D_2 = 4$. Note that although the function $\mu(x_1, x_2)$ produces a matrix rather than a vector of domain means, it can be vectorized in order to represent the limiting domain means as the vector $\boldsymbol{\mu}$. For each domain $d$, we generate its $N_d = N/D = 400$ elements by adding independent and normally distributed noise with mean 0 and variance $\sigma^2$ to the $\mu_d$. Once the elements of the population have been simulated, then the population domain means $\bar{\mathbf{y}}_U$ are computed. The population domain means used for simulations when $\sigma = 1$ are displayed in Figure 4.1. Observe that these domain means are reasonably (not strictly) monotone with respect to $x_1$ and $x_2$.

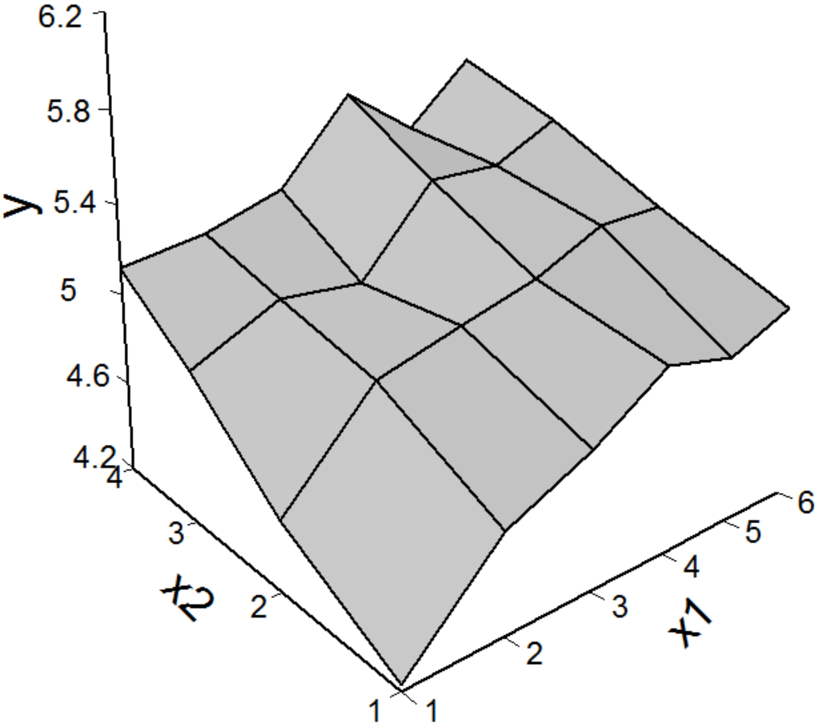


**Figure 4.1 Population domain means for simulations when $\sigma$ = 1.**

Samples are drawn from a stratified sampling design without replacement, with 4 strata that cut across the $D$ domains. Strata are constructed using an auxiliary variable $v$ that is correlated with the variable of interest $y$. The vector $v$ is created by adding independent standard normally distributed noise to $\sigma d / D$, for each element in domain $d$. Then, stratum membership is assigned by sorting the vector $v$ and creating 4 blocks of $N/4 = 2{,}400$ elements each based on the sorted $v$. To make the design informative, we sample $n = 480$ elements divided across strata in (60, 120, 120, 180). This probability sampling design is similar to the one described in Wu et al. (2016).

We consider 4 different scenarios obtained from the combination of two possible types of shape constraints and $\sigma = 1$ or 2. The first type of constraints assumes the population domain means are monotone increasing with respect to both $x_1$ and $x_2$ (double monotone), while the second type of constraints assumes monotonicity only with respect to $x_1$ (only $x_1$ monotone). For a fixed $\sigma$, the exact same population is considered for the two possible types of constraints. For each scenario, the unconstrained $\tilde{\mathbf{y}}_s$ and constrained $\tilde{\boldsymbol{\theta}}_s$ estimates are computed along with their linearization-based variance estimates (see (2.11)). Constrained estimates are computed using the CPA, and their variance estimates are computed by relying on the sample-selected set $J \in \mathcal{G}_s$. In addition, 95% Wald confidence intervals based on the normal distribution are constructed for both estimators.

To measure the precision of $\tilde{\mathbf{y}}_s$ and $\tilde{\boldsymbol{\theta}}_s$ as estimators of the population domain means $\bar{\mathbf{y}}_U$, we consider the Weighted Mean Squared Error (WMSE) given by

$$\text{WMSE}(\tilde{\boldsymbol{\varphi}}_s) = \mathbb{E}\left[(\tilde{\boldsymbol{\varphi}}_s - \bar{y}_U)^\top \mathbf{W}_U (\tilde{\boldsymbol{\varphi}}_s - \bar{y}_U)\right],$$

where $\tilde{\boldsymbol{\varphi}}_s$ could be either the unconstrained or constrained estimator and $\mathbf{W}_U$ is the diagonal matrix with elements $N_d / N$, $d = 1, \ldots, D$. The WMSE values are approximated by simulations as

$$\frac{1}{B}\sum_{b=1}^{B}(\tilde{\boldsymbol{\varphi}}_s^{(b)} - \bar{\mathbf{y}}_U)^\top \mathbf{W}_U (\tilde{\boldsymbol{\varphi}}_s^{(b)} - \bar{\mathbf{y}}_U),$$

where $B$ is the number of simulations, and $\tilde{\boldsymbol{\varphi}}_s^{(b)}$ is the estimator for the $b^{\text{th}}$ sample.

Simulation results are summarized in Figures 4.2 - 4.5, and are based on $R = 10{,}000$ replications. These display the 24 domains divided in groups of 6, where each group is assumed to be monotone. For the double monotone scenario, similar plots with groups of 4 monotone domains each can be also pictured. As illustrated in the fits of a single sample in these figures, it can be seen that the constrained estimates can be exactly equal to the unconstrained estimates for some domains. In those cases, their variance estimates are also equal. Overall, confidence intervals for the constrained estimator tend to be tighter in comparison with those for the unconstrained estimator. On average, the constrained estimator behaves slightly differently than the population domain means, due to the latter's non-strict monotonicity. As an advantage, the percentiles for the constrained estimator are narrower, demonstrating that the distribution of the proposed estimator is tighter than the distribution of the unconstrained estimator. For small values of $\sigma$, the unconstrained estimates are more likely to satisfy the assumed restrictions, which leads to small improvements on the constrained estimator over the unconstrained. In contrast, shape assumptions tend to be more severely violated in unconstrained estimates for larger values of $\sigma$, allowing the proposed estimator to gain much more efficiency on these cases. This latter property can be noted by observing that the constrained estimator percentile band gets farther away from the unconstrained estimator band as $\sigma$ increases.

In terms of variability, the constrained estimator has the smaller variance of the two estimators. Interestingly, it gets overestimated by its corresponding linearization-based variance estimate. In contrast, the variance estimate of the unconstrained estimator underestimates the true variance, which is a known and often observed drawback of linearization variances. Despite this difference, confidence intervals for both estimators demonstrate a similar good coverage rate when $\sigma = 1$, meanwhile such coverage gets slightly improved by the constrained estimator when $\sigma = 2$.

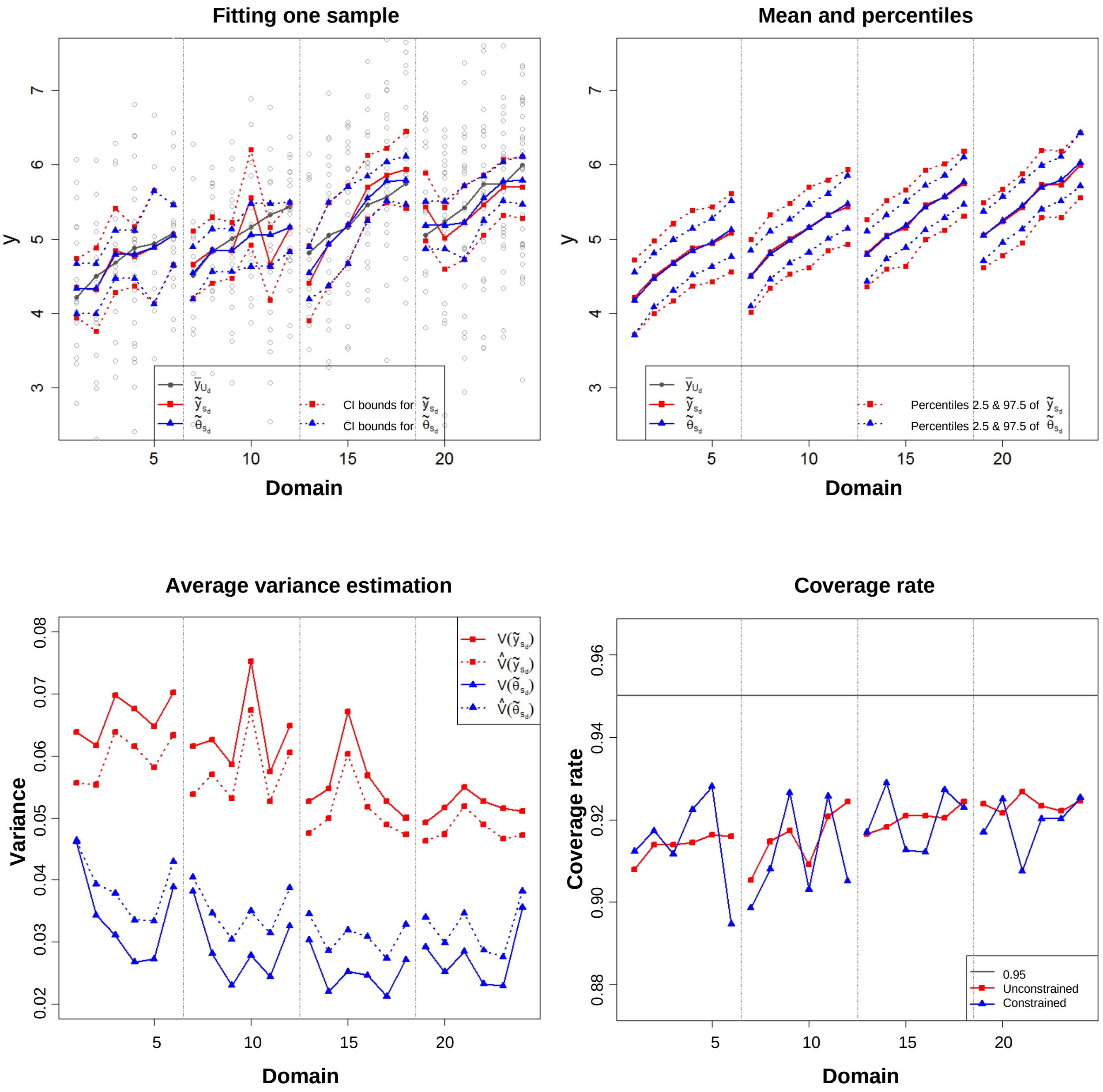


**Figure 4.2** **Plots of simulation results for the unconstrained and constrained estimators under the double monotone scenario with $\sigma = 1$. In the "Mean and percentiles" plot, $\bar{y}_{U_d}$ is hidden by $\tilde{y}_{s_d}$.**

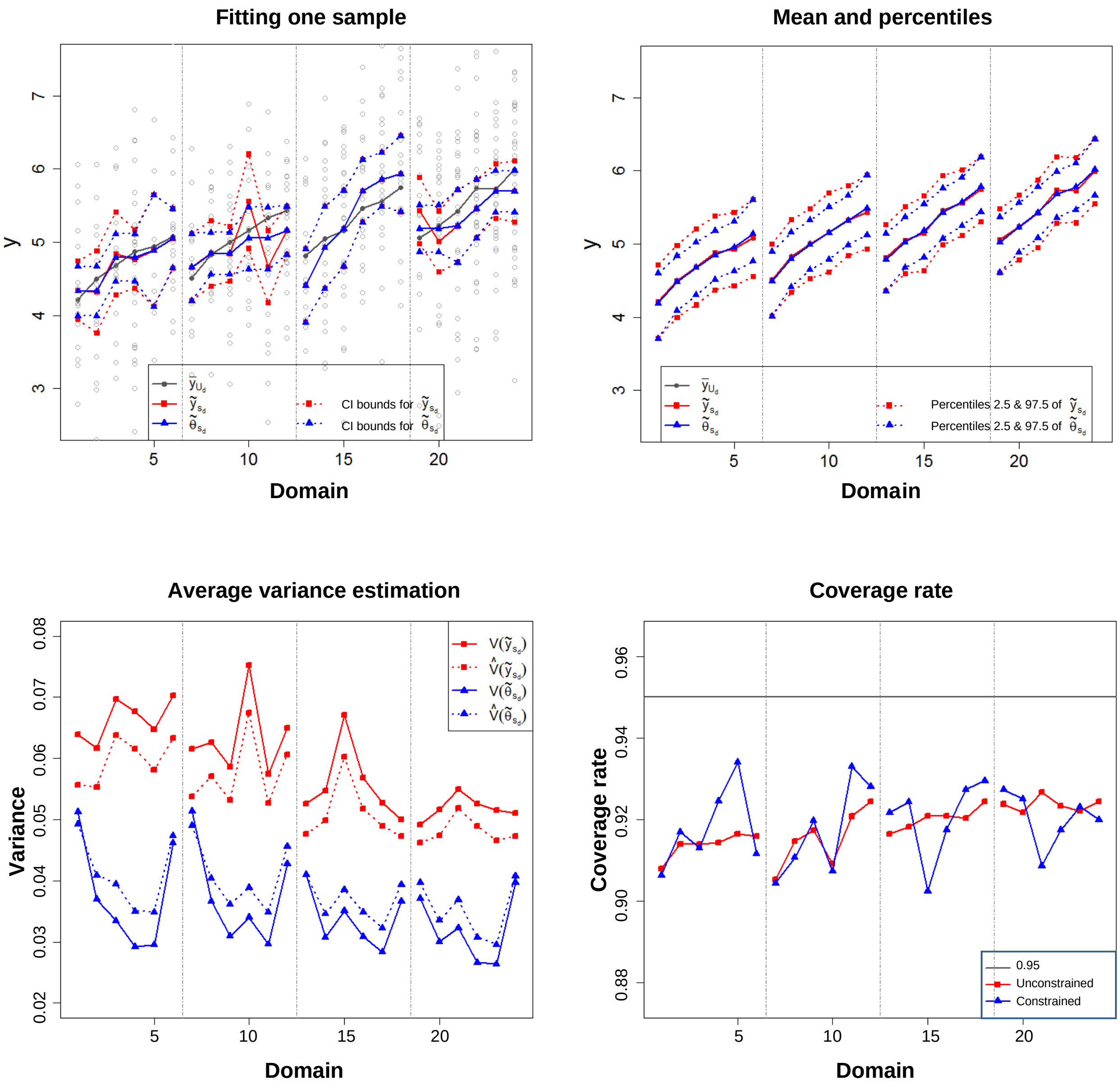


**Figure 4.3** **Plots of simulation results for the unconstrained and constrained estimators under the only $x_1$ monotone scenario with $\sigma = 1$. In the "Mean and percentiles" plot, $\bar{y}_{U_d}$ is hidden by $\tilde{y}_{s_d}$.**

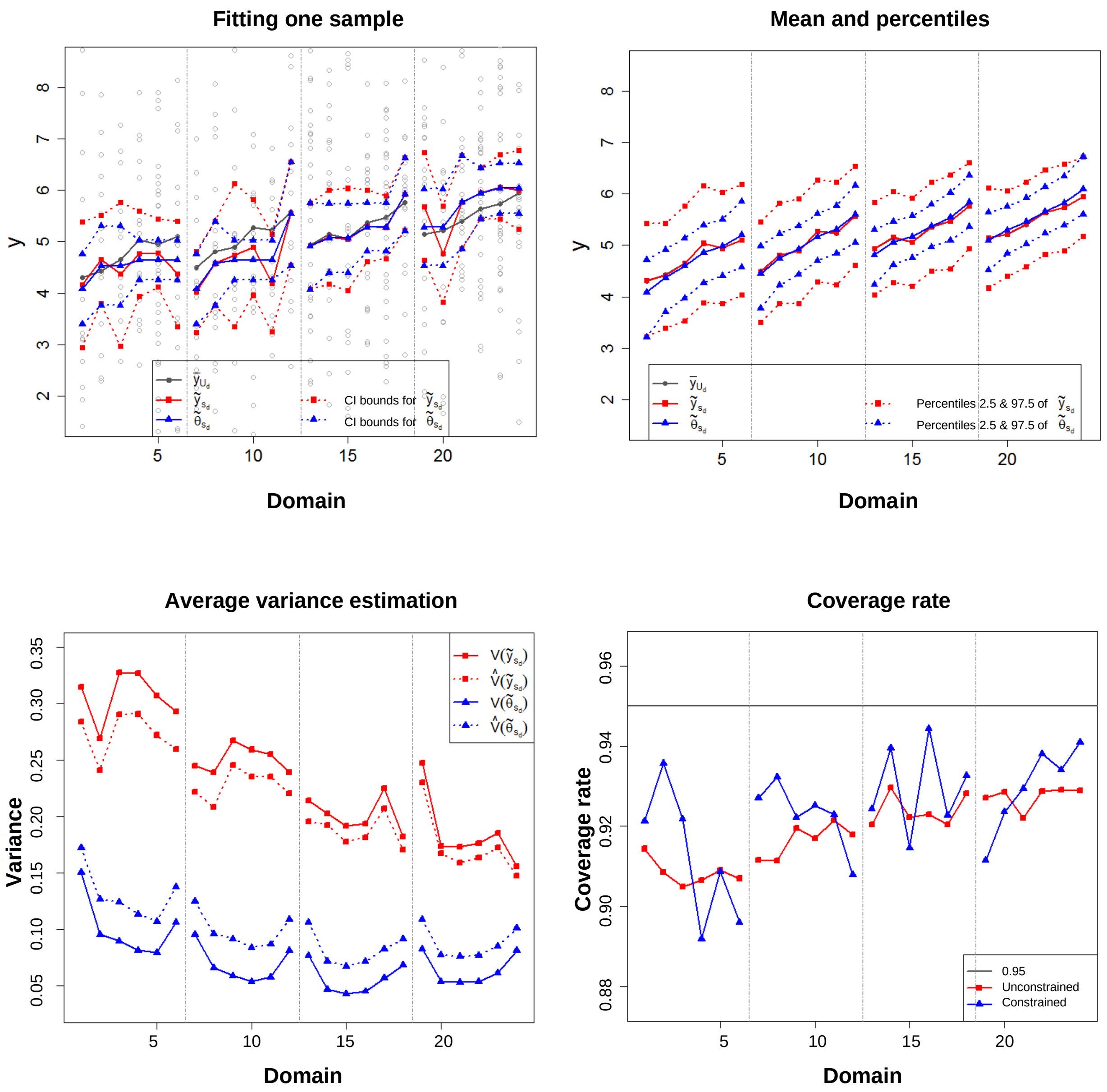


**Figure 4.4** **Plots of simulation results for the unconstrained and constrained estimators under the double monotone scenario with $\sigma = 2$. In the “Mean and percentiles” plot, $\bar{y}_{U_d}$ is hidden by $\tilde{y}_{s_d}$.**

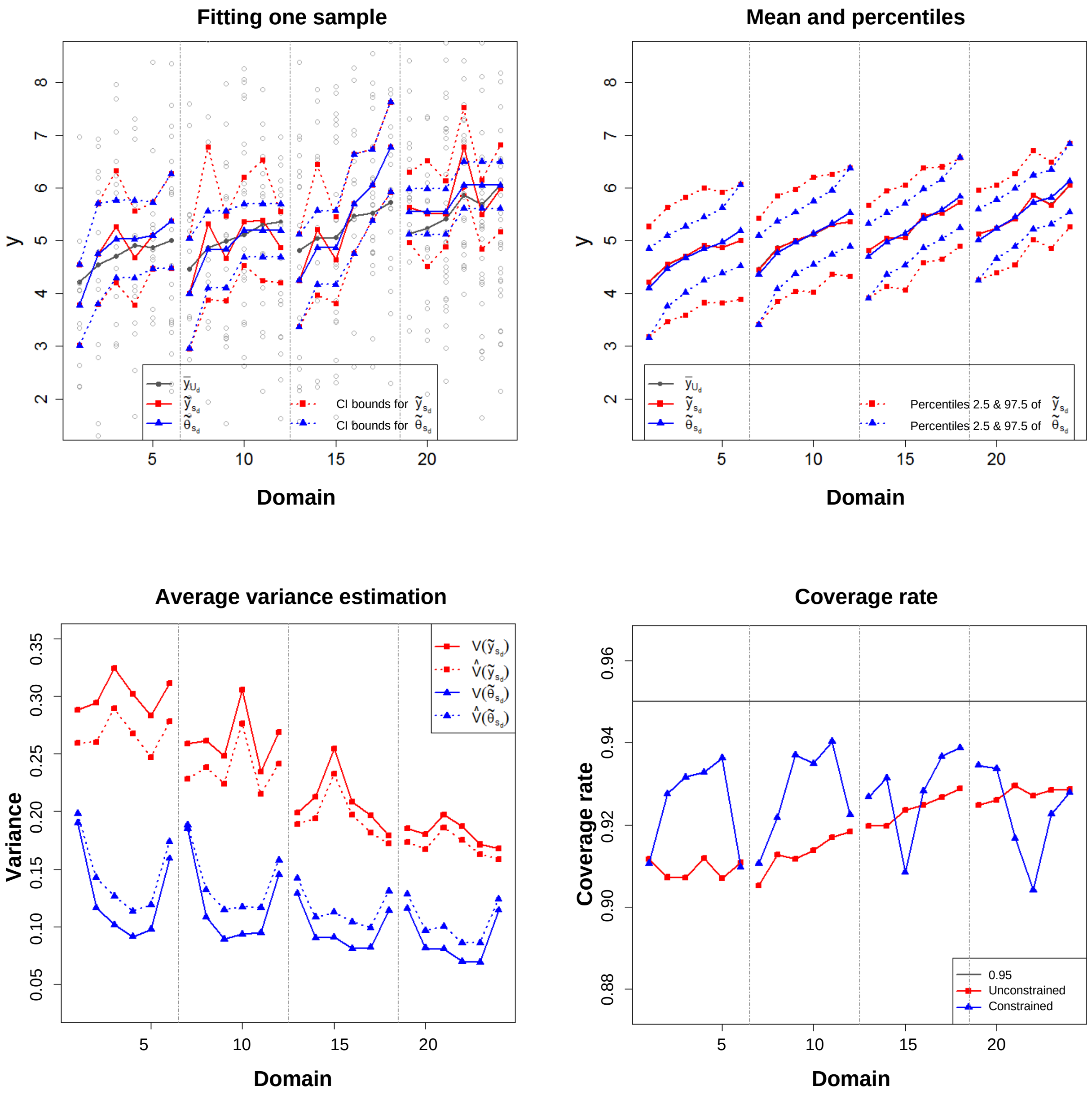


**Figure 4.5** **Plots of simulation results for the unconstrained and constrained estimators under the only $x_1$ monotone scenario with $\sigma = 2$. In the "Mean and percentiles" plot, $\bar{y}_{U_d}$ is hidden by $\tilde{y}_{s_d}$.**

Table 4.1 shows that the constrained estimator is more precise on average than the unconstrained estimator. The precision of the constrained estimator improves when the monotonicity with respect to the two variables is assumed, instead of only with respect to $x_1$. This is expected here, because the underlying surface is indeed doubly monotone, so that the estimator benefits from imposing the stronger constraint.

**Table 4.1**
**Empirical WMSE values**

| | Unconstrained | Only $x_1$ monotone | Double monotone |
|---|---|---|---|
| $\sigma = 1$ | 0.0593 | 0.0362 | 0.0298 |
| $\sigma = 2$ | 0.2384 | 0.1175 | 0.0832 |

## 4.2 Replication methods for variance estimation

In practice, it is common for large-scale surveys to use replication-based methods for variance estimation. Examples of such surveys are the last editions of the NHANES and the National Survey of College Graduates (NSCG). To study the performance of replication-based variance estimators under the proposed constrained methodology, we carry out simulation studies based on the delete-a-group Jackknife (DAGJK) variance estimator proposed by Kott (2001).

We perform replication-based simulation experiments using the setting described in Section 4.1. To compute the DAGJK variance estimator, we first randomly create $G$ equal-sized groups within each of the 4 strata. Then, for each replicate $g = 1, \ldots, G,$ we delete the $g^{\text{th}}$ group in each of the strata, adjust the remaining weights by $w_k^{(g)} = \left(\frac{G}{G-1}\right) w_k,$ where $w_k = \pi_k^{-1};$ and compute the replicate constrained estimate $\tilde{\boldsymbol{\theta}}_s^{(g)}$ using the adjusted weights. The DAGJK variance estimate of $\tilde{\theta}_{s_d}$, $\hat{V}_{\text{JK}}\left(\tilde{\theta}_{s_d}\right),$ is obtained by calculating

$$\hat{V}_{\text{JK}}\left(\tilde{\theta}_{s_d}\right) = \frac{G - 1}{G} \sum_{g=1}^{G} \left(\tilde{\theta}_{s_d}^{(g)} - \tilde{\theta}_{s_d}\right)^2.$$

A replication-based variance estimator of $\tilde{y}_{s_d}$ is obtained by substituting $\tilde{\boldsymbol{\theta}}_s$ by $\tilde{\mathbf{y}}_s$.

Our simulations consider only the double monotone scenario, with $\sigma = 1$ or 2, and $G = 10$, 20 or 30. The sample size is set to either $n = 480$ or $n = 960$, where the latter case is obtained by doubling the original sample size in each strata. Figures 4.6 - 4.9 contain simulation results based on 10,000 replications. In contrast to the behavior of the linearization-based variance estimates, it can be seen that the DAGJK estimates tend to overestimate the variance of the unconstrained estimator, as is often observed in practice. Both replication-based and linearization-based variance estimates of the constrained estimator overestimate the true variance, so that the results are more consistent across variance estimation methods. As the number of groups $G$ increases, DAGJK estimates tend to be greater, especially for small values of $\sigma$. Such increments on DAGJK estimates have the direct consequence of increasing the coverage rate as $G$ gets larger. In addition, the coverage rate for both estimators is improved (closer to

0.95) when the sample size is increased. Overall, it appears that replication variance estimation is a practical alternative to linearization.

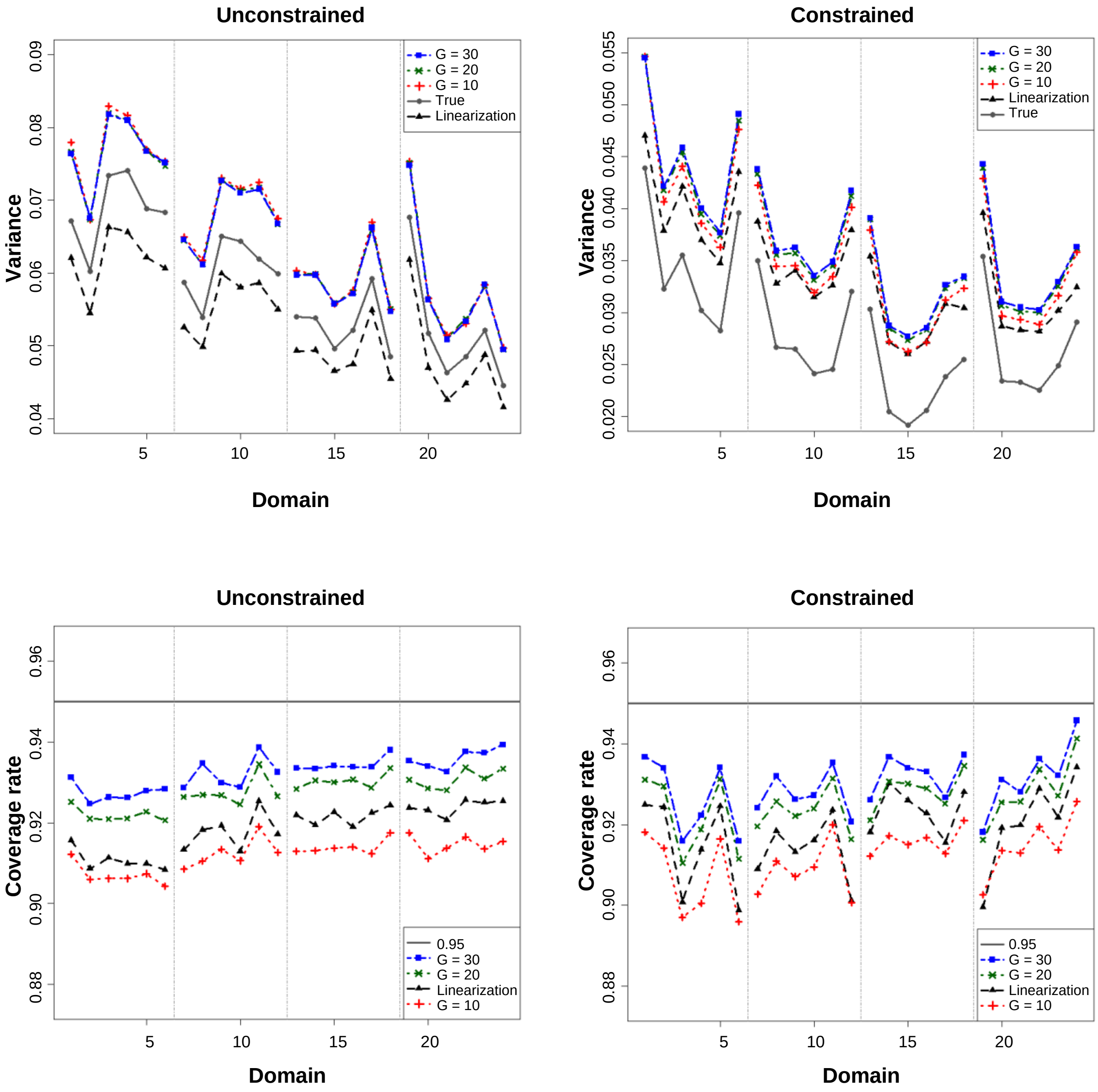


**Figure 4.6 Variance estimation (top) and coverage rate (bottom) simulation results based on linearization and DAGJK methods for the unconstrained (left) and constrained (right) estimators, under the double monotone scenario with $n_N = 480$ and $\sigma = 1$.**

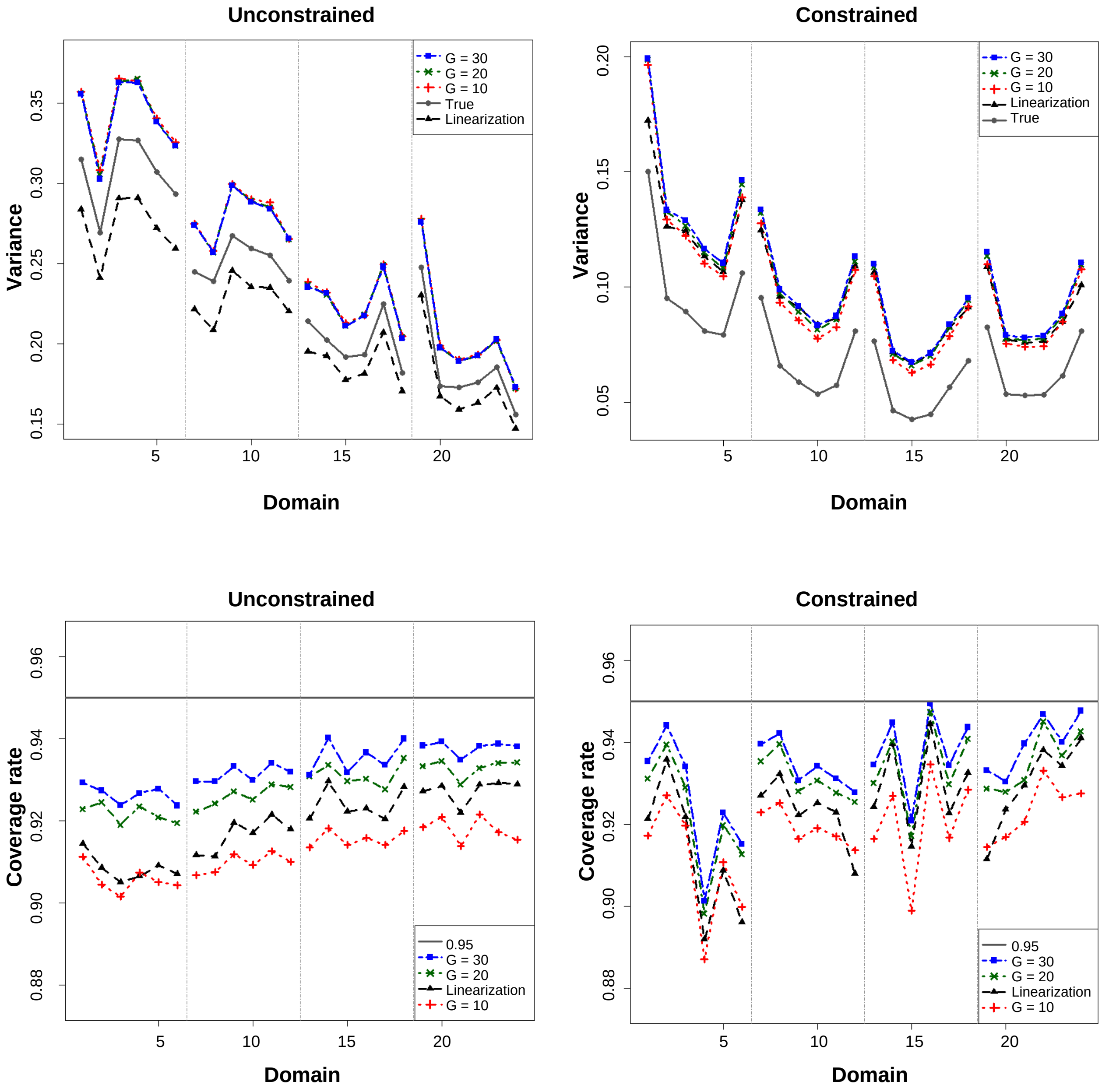


**Figure 4.7 Variance estimation (top) and coverage rate (bottom) simulation results based on linearization and DAGJK methods for the unconstrained (left) and constrained (right) estimators, under the double monotone scenario with $n_N = 480$ and $\sigma = 2$.**

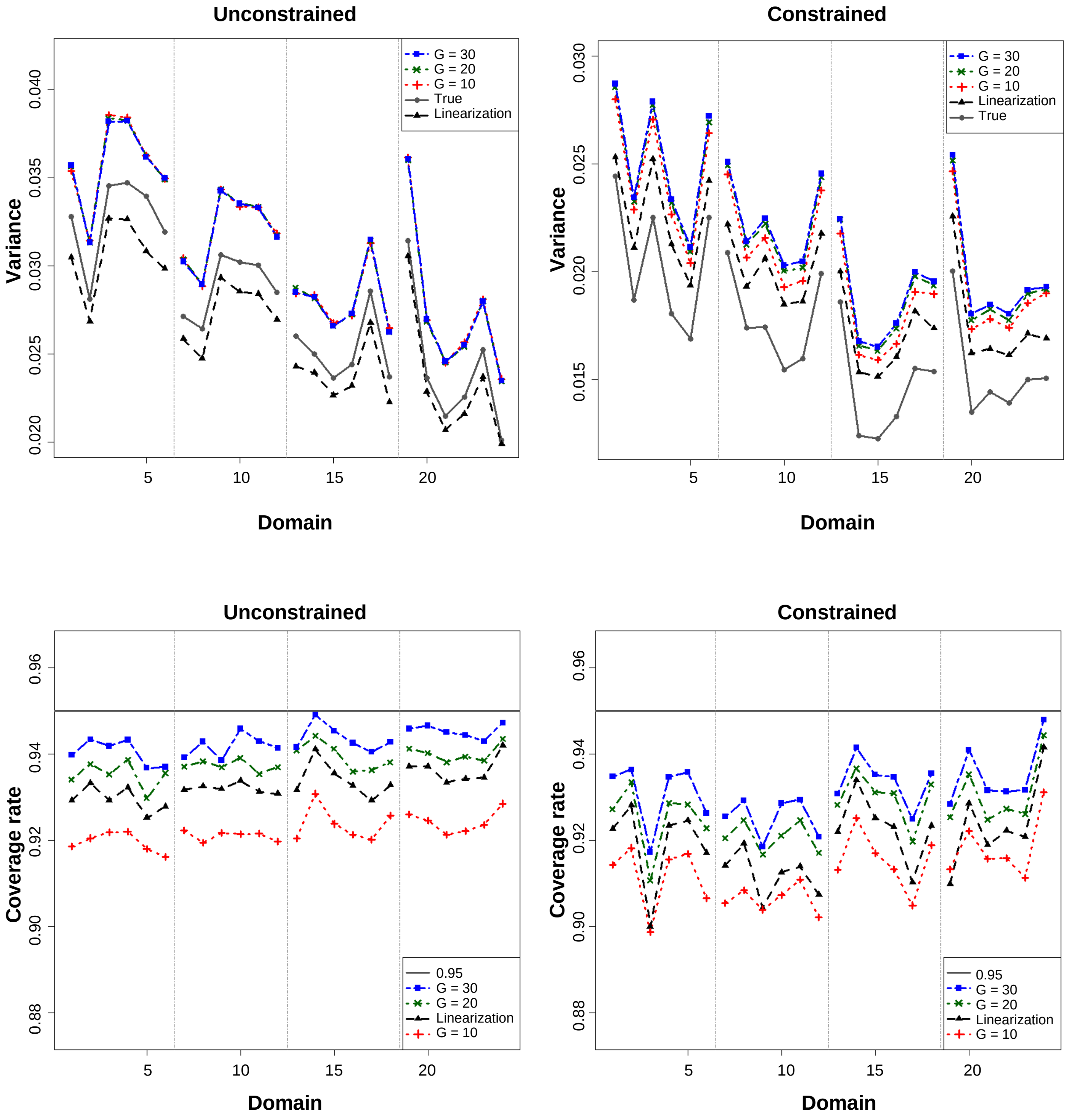


**Figure 4.8** **Variance estimation (top) and coverage rate (bottom) simulation results based on linearization and DAGJK methods for the unconstrained (left) and constrained (right) estimators, under the double monotone scenario with $n_N = 960$ and $\sigma = 1$.**

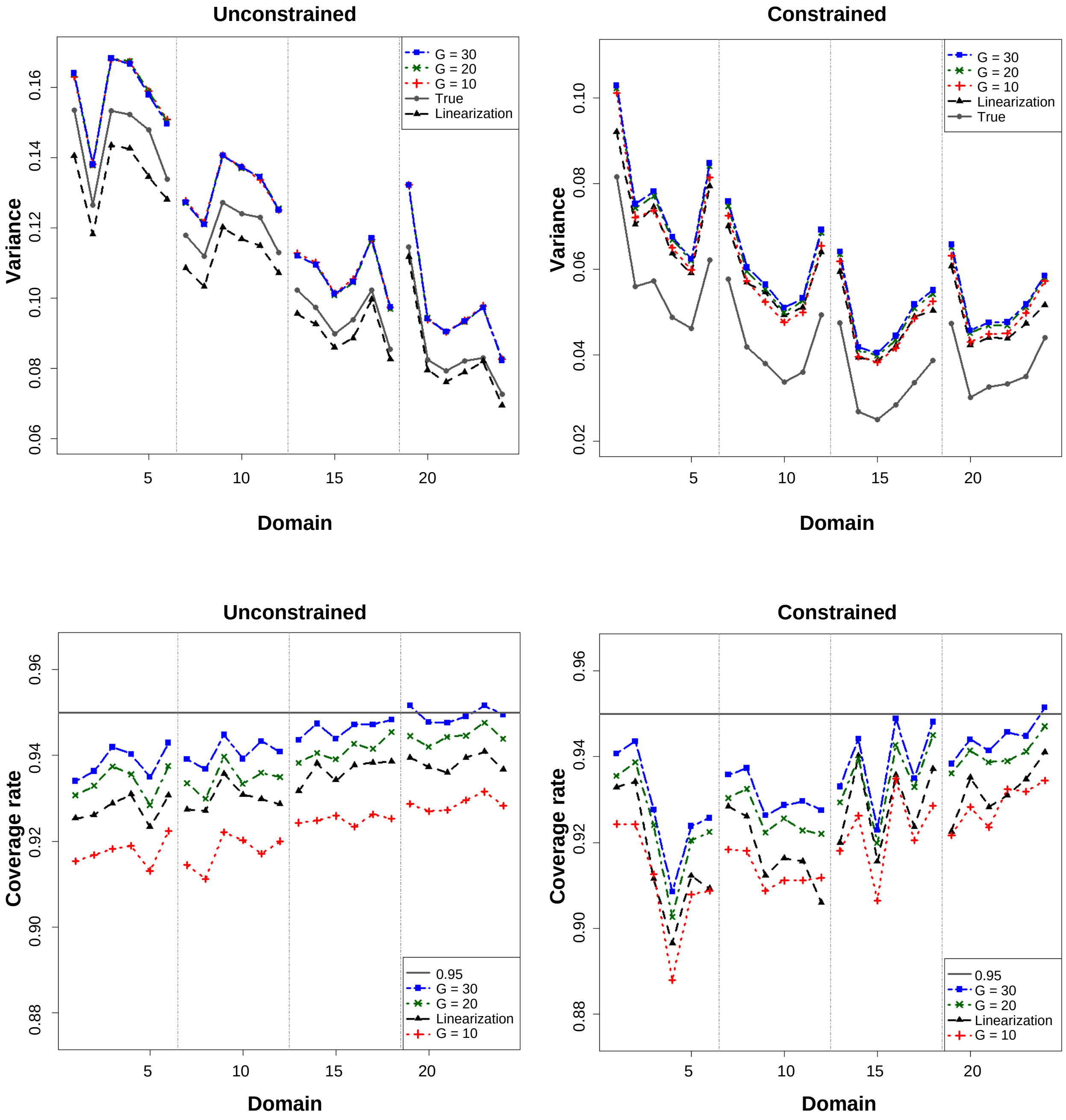


**Figure 4.9 Variance estimation (top) and coverage rate (bottom) simulation results based on linearization and DAGJK methods for the unconstrained (left) and constrained (right) estimators, under the double monotone scenario with $n_N = 960$ and $\sigma = 2$.**

# 5 Application of constrained estimator to NSCG

To demonstrate the utility of the proposed constrained methodology in real survey data, we consider the 2015 National Survey of College Graduates (NSCG), which is sponsored by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation, and is conducted by the U.S. Census Bureau. The 2015 NSCG data and documentation are available on the NSF website (www.nsf.gov/statistics/srvygrads). The purpose of the NSCG is to provide data on the characteristics of U.S. college graduates, with particular focus on those in the science and engineering workforce.

We consider the total earned income before deductions in previous year (2014) to be the variable of interest (denoted by EARN). To avoid the high skewness of this variable, a log transformation is performed. Moreover, we take into account only those who reported a positive earning amount. A total of 76,389 observations was considered in our analysis. In addition, 252 domains are considered. These are determined by the cross-classification of four predictor variables. These variables and their assumed constraints are as follows:

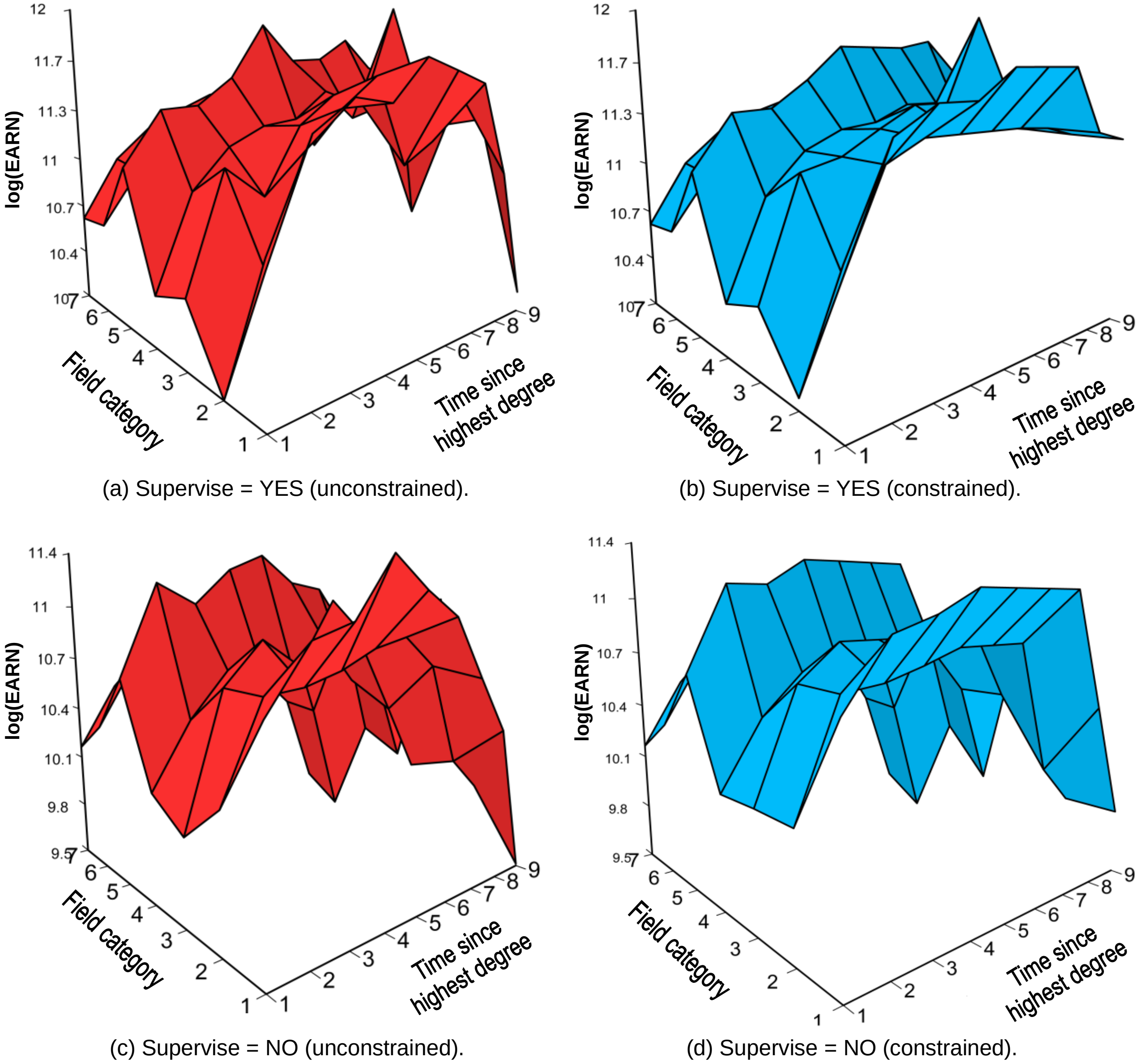
- *Time since highest degree*. This variable defines the year of award of highest degree. The period from 2015 to 1959 is divided into 9 categories, where the first 8 categories (denoted by 1-8) are of 6 years each, and the last category (denoted by 9) is of 9 years. *Constraint:* given the other predictors, the average total earned income increases with respect to the time since highest degree from year category 1 to 7. No assumption is made with respect to categories 8 and 9, as those people are likely to be retired (at least 42 years since their highest degree).
- *Field of study*. This nominal variable defines the field of study for highest degree, based on a major group categorization provided within the 2015 NSCG. The 7 categories for this variable are:

  1: Computer and mathematical sciences,
  2: Biological, agricultural and environmental life sciences,
  3: Physical and related sciences,
  4: Social and related sciences,
  5: Engineering,
  6: S&E-related fields,
  7: Non-S&E fields.

  *Constraint:* given the other predictors, the average total earned income for each of the fields 2 and 4 is less than for the fields 1, 3 and 5. No assumption is made with respect to categories 6 and 7, as they cover many fields for which a reasonable order restriction might be complicated to impose.
- *Postgrad.* This binary variable defines whether the highest degree is at the postgraduate level (YES) or at the Bachelor's level (NO). *Constraint:* given the other predictors, the average total earned income is higher for those with postgraduate studies.

- *Supervise.* This binary variable defines whether supervising others is a responsibility in the principal job (YES) or not (NO). *Constraint:* given the other predictors, the average total earned income is higher for those who supervise others in their principal job.

Figures 5.1 and 5.2 show the unconstrained and constrained estimates for each of the four groups obtained from the cross-classification of the Postgrad and Supervise binary variables. Note that since the assumed constraints constitute a partial ordering, then the constrained estimates are obtained by pooling domains. These figures show that the constrained estimator has a smoother behavior than the unconstrained. Moreover, it tends to correct for the some of the "spikes" produced by the unconstrained estimator, which are usually a consequence of a very small sample size.



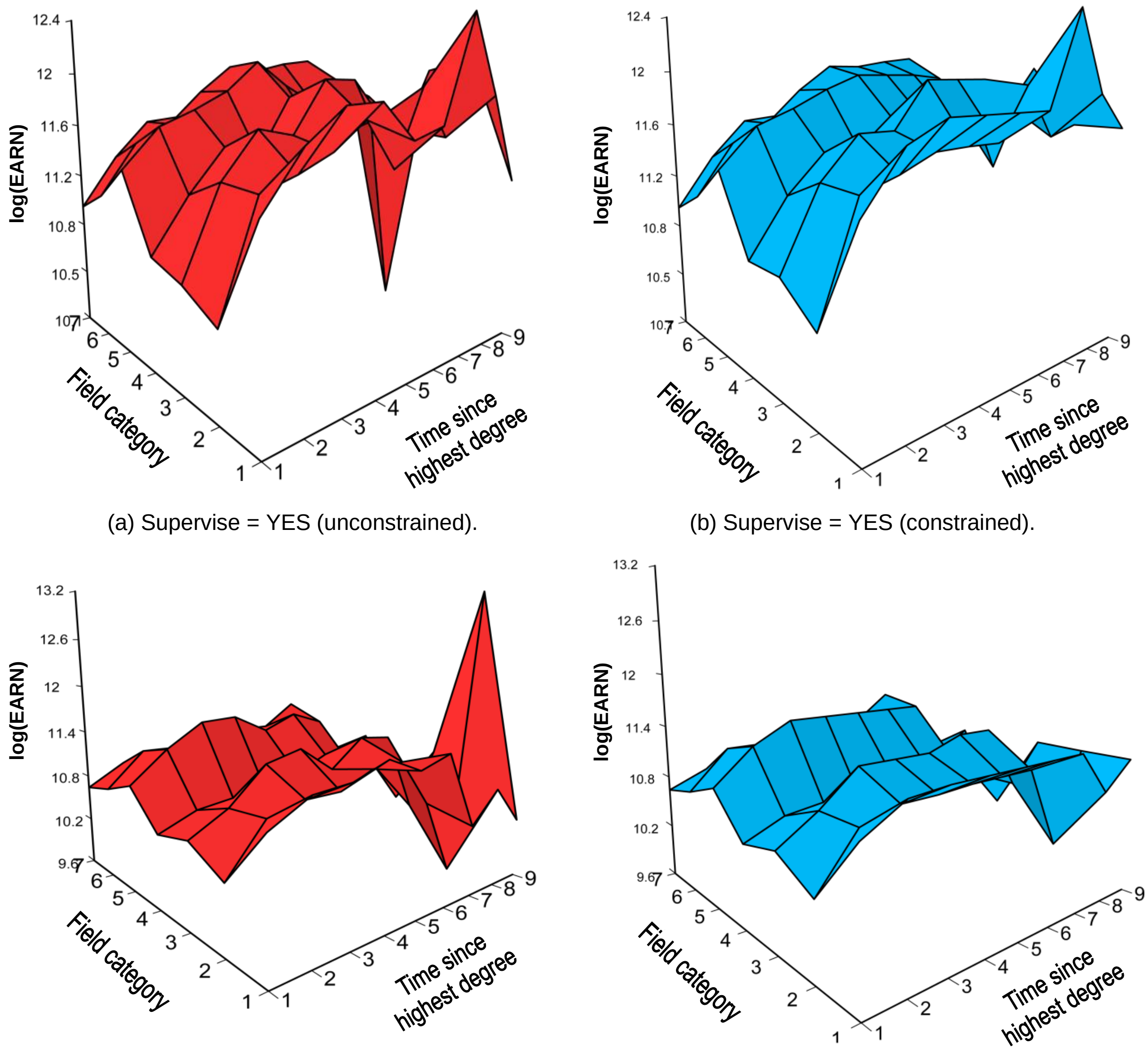
(a) Supervise = YES (unconstrained).

(b) Supervise = YES (constrained).

(c) Supervise = NO (unconstrained).

(d) Supervise = NO (constrained).

**Figure 5.1 Unconstrained (left) and constrained (right) domain mean estimates for the 2015 NSCG data, given that Postgrad = NO is fixed.**



(a) Supervise = YES (unconstrained).

(b) Supervise = YES (constrained).

(c) Supervise = NO (unconstrained).

(d) Supervise = NO (constrained).

**Figure 5.2 Unconstrained (left) and constrained (right) domain mean estimates for the 2015 NSCG data, given that Postgrad = YES is fixed.**

Standard errors for both unconstrained and constrained estimates are computed using the 2015 NSCG replicate weights, which are based on successive difference replication method (Opsomer, Breidt, White and Li, 2016). The replicate weights and adjustment factors were provided by the Program Director of the Human Resources Statistics Program from the NCSES and are available upon request.

Figure 5.3 displays the ratio of these estimates for each of the 252 domains. In the vast majority of cases, the standard error estimates of the proposed estimator are lower than those for the unconstrained estimator, with improvements of as much as 7 times smaller. However, there are some cases where the opposite behavior occurs. These are investigated in Figure 5.4, which shows plots of two different domain "slices": one with respect to the Time since highest degree variable and other with respect to Field category. These plots include unconstrained and constrained estimates, Wald confidence intervals and sample sizes. Each of these two slices contain one of the two domains that can be easily identified in Figure 5.3 to have the smallest ratios.

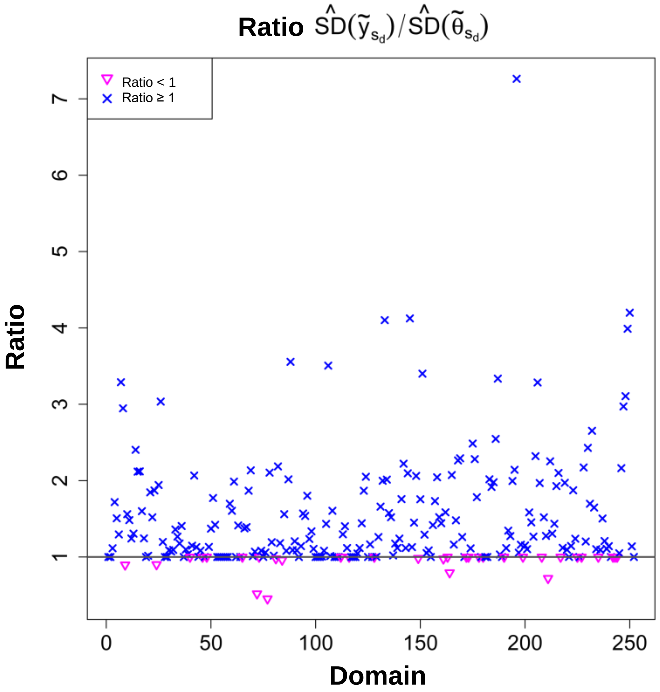
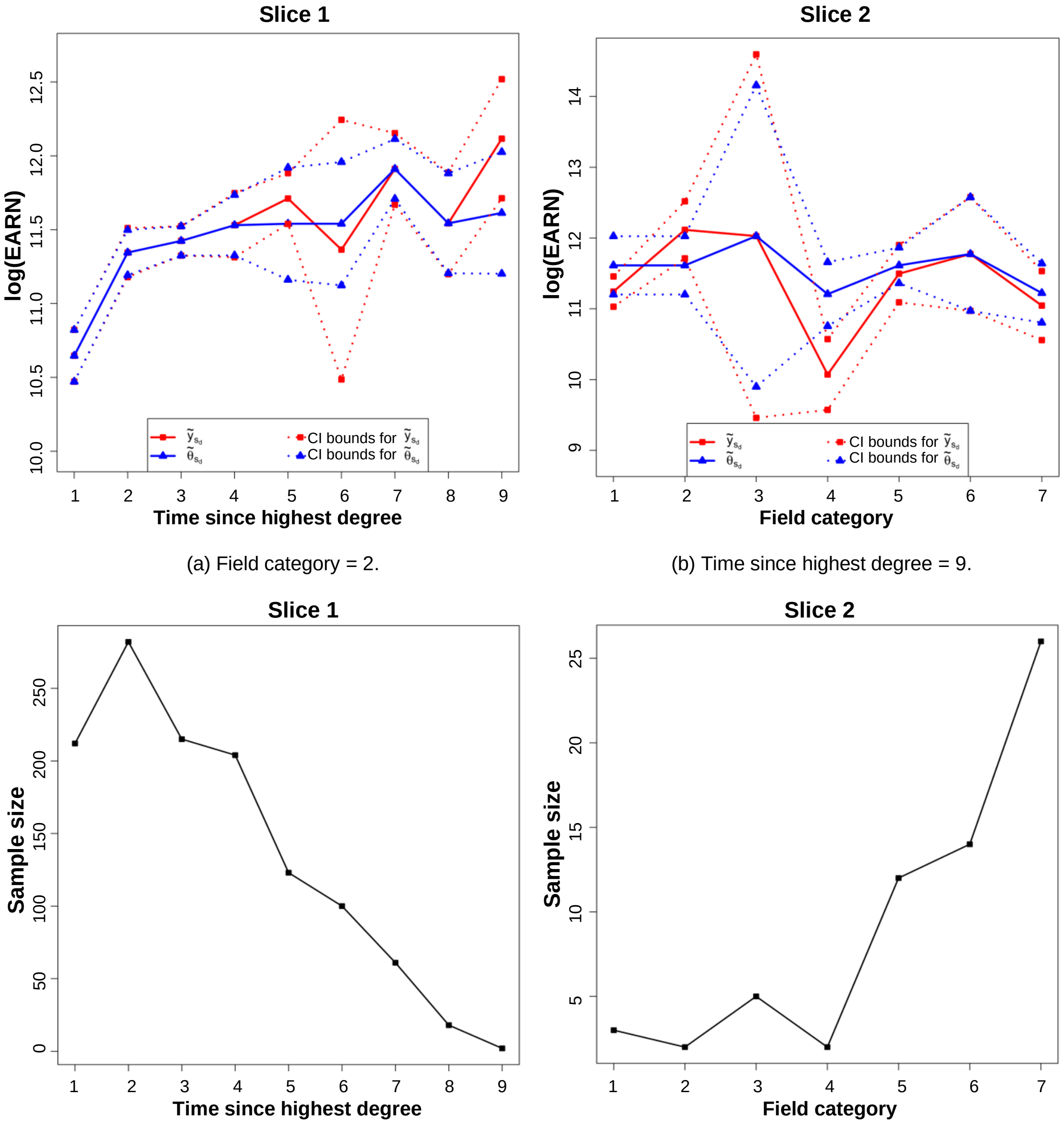


**Figure 5.3 Ratio of the estimated standard errors of unconstrained estimates over those for constrained estimates for the 2015 NSCG data.**

The first of these domains is displayed in Figure 5.4(a) and 5.4(c), indexed by 5. The unconstrained estimates for the domains indexed by 5 and 6 violate the monotonicity assumption, and thus, are being pooled to obtain the constrained estimates (additional pooling with domains in other "slices" is also occurring, but not visible in this plot). As can be seen in Figure 5.4(a), the confidence interval is narrower for the unconstrained estimates. However, the estimated standard error of the unconstrained estimator of domain 6 is very large, and pooling with domain 5 greatly stabilizes both the estimator and the estimated standard errors for that domain. Figure 5.4(c) shows that the samples sizes on these domains are reasonably large at approximately 100 observations each, implying that the noticed monotonicity violation might be in fact true in the population. The final decision on the balance between the improved stability of some domains with the potential for bias due to incorrect constraints would need to be carefully evaluated.

The second domain where unconstrained estimates produce smaller standard deviation estimates is displayed in Figure 5.4(b) and 5.4(d), indexed by 1. Here, this domain is being pooled with its neighboring domain to obtain the constrained estimate. However, as these two domains have very low sample sizes, the unconstrained estimates might be considered as unreliable, so that their estimated standard errors are not a good indication of their precision. The constrained estimator appears to be preferred here because of the increase in the effective cell size.



(a) Field category = 2.

(b) Time since highest degree = 9.

(c) Field category = 2.

(d) Time since highest degree = 9.

**Figure 5.4 Unconstrained and constrained estimates with Wald confidence intervals (top) and sample sizes (bottom) for the 2015 NSCG data, given that Postgrad = YES and Supervise = YES.**

# 6 Conclusions

We have proposed a general methodology to estimate domain means which makes it possible incorporate natural restrictions between domains into design-based estimation. It was shown to improve estimation and inference, especially on small domains. As this new methodology covers a broad range of shape assumptions beyond univariate monotonicity, it aims to jointly take advantage of several types of qualitative information that arises naturally for survey data. Additional shapes that may be imposed include convexity or log-concavity; the latter might be imposed if the population domain means are believed to be increasing and then decreasing over a set of domains. Future work by the authors will include a “relaxed monotone” estimator to be used when the population domain means are “roughly” monotone in some sequence of domains. For the relaxed monotone estimator, a type of moving average over the domains is used to implement the constraints, allowing the estimator to have some departures from monotonicity.

We also proposed a design-based variance estimation method of the estimator, which only requires knowledge of the sample-specific constraint set. Replication-based methods are shown to behave similarly. From the computational side, the estimator is based on the Cone Projection Algorithm which is efficiently implemented in the package `coneproj` and freely available. In the important practical case of partial ordering, the constrained estimator is equivalent to a pooling of neighboring domains, so that once the constraint set is identified by CPA, subsequent computations of estimators and variance estimators can be done directly using traditional design-based estimation for the relevant domains.

An important practical issue, as illustrated in the NSCG analysis in Section 5, is the determination of when the imposed constraint might not be valid for a particular survey application. Recently, Oliva-Aviles, Meyer and Opsomer (2019) proposed the sample-based Cone Information Criterion as a criterion to choose between the constrained and unconstrained fits for the estimator of Wu et al. (2016). That approach is generalizable to the setting considered here, and is currently under development.

# Appendix

The first part of this appendix contains lemmas used to obtain the theoretical results discussed in this paper. Proofs of the theorems are included at the end of this appendix.

**Lemma 1.** *If a non-zero vector can be written as the positive linear combination of linearly dependent non-zero vectors, then it can be expressed as the positive linear combination of a linearly independent subset of these.*

*Proof.* Let $\mathbf{v}$ be a non-zero vector such that it can be written as $\mathbf{v} = \sum_{i=1}^{k} a_i \boldsymbol{\ell}_i$ where $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_k$ are non-zero vectors and $a_i > 0$ for $i = 1, 2, \ldots, k$. If this set of vectors is not linearly independent, then there exist constants $b_1, \ldots, b_k$, not all zero, such that $\sum_{i=1}^{k} b_i \boldsymbol{\ell}_i = \mathbf{0}$, and for any $c \in \mathbf{R}$, $\mathbf{v} = \sum_{i=1}^{k} (a_i + cb_i) \boldsymbol{\ell}_i$. Let $c = -\min_{i:\, b_i \neq 0} a_i / b_i$; then $a_i + cb_i \geq 0$ for $i = 1, \ldots, k$ but for at least one $i$, $a_i + cb_i = 0$. Then we have written $\mathbf{v}$ as a positive linear combination of a proper subset of the vectors. If this subset is still linearly dependent, the process can be repeated.

**Lemma 2.** *If $\mathbf{A}$ is a $m \times D$ irreducible matrix and $\mathbf{B}$ is a $D \times D$ nonsingular matrix, then $\tilde{\mathbf{A}} = \mathbf{AB}$ is also irreducible.*

*Proof.* Suppose $\tilde{\mathbf{A}}^\top \mathbf{c} = \mathbf{0}$ for some $\mathbf{c} \in \mathbb{R}^m, \mathbf{c} \geq \mathbf{0}$. Then $\mathbf{B}^\top\mathbf{A}^\top\mathbf{c} = \mathbf{0}$ implies that $\mathbf{A}^\top\mathbf{c} = \mathbf{0}$ by the non-singularity of $\mathbf{B}$. Because $\mathbf{A}$ is irreducible, we must have $\mathbf{c} = \mathbf{0}$, so the origin is not a positive linear combination of rows of $\tilde{\mathbf{A}}$. Next, suppose that one of the rows of $\tilde{\mathbf{A}}$ is a positive linear combination of other rows of $\tilde{\mathbf{A}}$. This means we can write $\tilde{\mathbf{A}}^\top\mathbf{b} = \mathbf{0}$, where $b_j = -1$ for some $j \in \{1, \ldots, m\}$ and $b_i \geq 0, i \neq j$. But $\tilde{\mathbf{A}}^\top\mathbf{b} = \mathbf{0}$ implies that $\mathbf{B}^\top\mathbf{A}^\top\mathbf{b} = \mathbf{0}$ implies that $\mathbf{A}^\top\mathbf{b} = \mathbf{0}$ by the non-singularity of $\mathbf{B}$. We can't have $\mathbf{A}^\top\mathbf{b} = \mathbf{0}$ for this $\mathbf{b}$, so we can't have a row of $\tilde{\mathbf{A}}$ is a positive linear combination of other rows of $\tilde{\mathbf{A}}$. Therefore, $\tilde{\mathbf{A}}$ is irreducible.

**Lemma 3.** *Let $\mathbf{A}$ be a $m \times D$ matrix. Also, let $\mathbf{S}_1$ and $\mathbf{S}_2$ be $D \times D$ diagonal matrices with nonzero elements on the diagonal. For any set $J \subseteq \{1, 2, \ldots, m\}$, denote $V_{i,J}$ to be the set of vectors in rows $J$ of $\mathbf{A}_i = \mathbf{AS}_i, i = 1, 2$. Then, for any $J^* \subseteq J$,*

$$\mathcal{L}\left(V_{1,J^*}\right) = \mathcal{L}\left(V_{1,J}\right) \Leftrightarrow \mathcal{L}\left(V_{2,J^*}\right) = \mathcal{L}\left(V_{2,J}\right).$$

*Proof.* Let $\mathbf{A}_{i,J} = \mathbf{A}_J\mathbf{S}_i, i = 1, 2$; where $\mathbf{A}_J$ denotes the submatrix of $\mathbf{A}$ that contains the rows in positions $J$. First, assume that $\mathcal{L}\left(V_{1,J^*}\right) = \mathcal{L}\left(V_{1,J}\right)$. Since $J^* \subseteq J$, it is straightforward to see that $\mathcal{L}\left(V_{2,J^*}\right) \subseteq \mathcal{L}\left(V_{2,J}\right)$. Now, consider any $\mathbf{v} \in \mathcal{L}\left(V_{2,J}\right)$ so that $\mathbf{v} = \mathbf{A}_{2,J}^\top \mathbf{a} = \mathbf{S}_2\mathbf{A}_J^\top \mathbf{a}$ for some vector $\mathbf{a}$. Then, we have $\mathbf{S}_1\mathbf{S}_2^{-1}\mathbf{v} = \mathbf{S}_1\mathbf{A}_J^\top \mathbf{a} \in \mathcal{L}\left(V_{1,J}\right)$. By assumption, there exists a vector $\mathbf{b}$ such that $\mathbf{S}_1\mathbf{S}_2^{-1}\mathbf{v} = \mathbf{S}_1\mathbf{A}_{J^*}^\top \mathbf{b}$. Therefore, $\mathbf{v} = \mathbf{S}_2\mathbf{A}_{J^*}^\top \mathbf{b} \in \mathcal{L}\left(V_{2,J^*}\right)$. Thus, $\mathcal{L}\left(V_{2,J}\right) \subseteq \mathcal{L}\left(V_{2,J^*}\right)$. Analogously, it follows that $\mathcal{L}\left(V_{2,J^*}\right) = \mathcal{L}\left(V_{2,J}\right)$ implies $\mathcal{L}\left(V_{1,J^*}\right) = \mathcal{L}\left(V_{1,J}\right)$.

**Lemma 4.** *Under Assumptions A1-A5, the following statements hold:*

*(i)* *The $N^{-1}\hat{t}_d$ are uniformly bounded.*

*(ii)* *The $N^{-1}\hat{N}_d$ are uniformly bounded above and uniformly bounded away from zero.*

*(iii)* $var\left(N^{-1}\hat{t}_d\right) = O\left(n^{-1}\right)$ *and* $var\left(N^{-1}\hat{N}_d\right) = O\left(n^{-1}\right)$.

*(iv)* $\mathbb{E}\left[\left(N^{-1}\hat{t}_d - r_d\mu_d\right)^2\right] = O\left(n^{-1}\right)$ *and* $\mathbb{E}\left[\left(N^{-1}\hat{N}_d - r_d\right)^2\right] = O\left(n^{-1}\right)$.

*Proof.*

(i) Note that

$$\frac{|\hat{t}_d|}{N} = \left|\frac{\sum_{k \in s_d} y_k / \pi_k}{N}\right| \leq \frac{\sum_{k \in U} |y_k|}{\lambda N}$$

which does not depend on $s$, and is bounded independently of $N$ by Assumption A2.

(ii) From Assumptions A4 and A5, note that

$$\frac{\varepsilon n}{DN} \le \frac{n_d}{N} \le \frac{\hat{N}_d}{N} = N^{-1} \sum_{k \in s_d} 1/\pi_k \le \lambda^{-1} N^{-1} N_d \le \lambda^{-1},$$

where both lower and upper bounds do not depend on $s$, and are bounded for all $N$ by Assumptions A1 and A4.

(iii) Note that

$$n \operatorname{var}\left(N^{-1} \hat{t}_d\right) = n \operatorname{var}\left(N^{-1} \sum_{k \in s_d} y_k / \pi_k\right) \le \frac{\sum_{k \in U_d} y_k^2}{\lambda^2 N} \left(\frac{n}{N} + n \max_{k,\, l \in U_d:\ k \ne l} |\Delta_{kl}|\right)$$

which is bounded by Assumptions A2, A4 and A5. Setting $y_k \equiv 1$ and following an analogous argument, it can be shown that $n \operatorname{var}\left(N^{-1} \hat{N}_d\right) = O(1)$.

(iv) Since

$$\mathbb{E}\left[\left(N^{-1} \hat{t}_d - r_d \mu_d\right)^2\right] = \operatorname{var}\left(N^{-1} \hat{t}_d\right) + \left(\frac{N_d}{N} \bar{y}_{U_d} - r_d \mu_d\right)^2,$$

Assumption A3 and (iii) lead to the desired conclusion. Analogously, we find

$$\mathbb{E}\left[\left(N^{-1} \hat{N}_d - r_d\right)^2\right] = O\left(n^{-1}\right).$$

*Proof of Theorem 1.* First, suppose that $\Pi\left(\mathbf{z} | \Omega^0\right) = \Pi\left(\mathbf{z} | \mathcal{L}(V_J)\right) = \mathbf{0}$. In that case, any subset $J^* \subset J$ such that $V_J$ is linearly independent will satisfy $\Pi\left(\mathbf{z} | \mathcal{L}(V_{J^*})\right) = \mathbf{0} \in \bar{\mathcal{F}}_{J^*}$. Hence, it is enough to choose $J^* \subset J$ such that $V_{J^*}$ is linearly independent and spans $\mathcal{L}(V_J)$. Now, suppose that $\Pi\left(\mathbf{z} | \Omega^0\right) \ne \mathbf{0}$. Since $\Pi\left(\mathbf{z} | \Omega^0\right) = \Pi\left(\mathbf{z} | \mathcal{L}(V_J)\right) \in \bar{\mathcal{F}}_J, \Pi\left(\mathbf{z} | \mathcal{L}(V_J)\right)$ can be written as the positive linear combination of vectors $\boldsymbol{\gamma}_j$, $j \in J$. Moreover, $\left\langle \mathbf{z} - \Pi\left(\mathbf{z} | \mathcal{L}(V_J)\right), \boldsymbol{\gamma}_j \right\rangle = 0$ for $j \in J$. From Lemma 1, there exists $J_0 \subset J$ such that $V_{J_0}$ is linearly independent and $\Pi\left(\mathbf{z} | \mathcal{L}(V_J)\right)$ can be written as a positive linear combination of the vectors in $V_{J_0}$, which implies that $\Pi\left(\mathbf{z} | \mathcal{L}(V_J)\right) \in \bar{\mathcal{F}}_{J_0}$. In addition, since $\left\langle \mathbf{z} - \Pi\left(\mathbf{z} | \mathcal{L}(V_J)\right), \boldsymbol{\gamma}_j \right\rangle = 0$ for $j \in J_0$, $\Pi\left(\mathbf{z} | \mathcal{L}\left(V_{J_0}\right)\right) = \Pi\left(\mathbf{z} | \mathcal{L}(V_J)\right)$. Thus, $\Pi\left(\mathbf{z} | \Omega^0\right) = \Pi\left(\mathbf{z} | \mathcal{L}\left(V_{J_0}\right)\right)$. If $\mathcal{L}\left(V_{J_0}\right) = \mathcal{L}(V_J)$ then $J^* = J_0$ satifies all required conditions. Now, assume that $\mathcal{L}\left(V_{J_0}\right) \subset \mathcal{L}(V_J)$. The fact that $\Pi\left(\mathbf{z} | \mathcal{L}\left(V_{J_0}\right)\right) = \Pi\left(\mathbf{z} | \mathcal{L}(V_J)\right)$ implies that $\Pi\left(\mathbf{z} | \mathcal{L}\left(V_{J_1}\right)\right) = \Pi\left(\mathbf{z} | \mathcal{L}\left(V_{J_0}\right)\right)$ for any set $J_1$ such that $J_0 \subseteq J_1 \subseteq J$. Further, since $\Pi\left(\mathbf{z} | \mathcal{L}\left(V_{J_0}\right)\right) \in \bar{\mathcal{F}}_{J_0}$ then $\Pi\left(\mathbf{z} | \mathcal{L}\left(V_{J_1}\right)\right) \in \bar{\mathcal{F}}_{J_1}$. Thus, it is enough to choose the set $J^*$ such that $J_0 \subset J^* \subset J$ and $V_{J^*}$ is a linearly independent set that spans $\mathcal{L}(V_J)$.

*Proof of Theorem 2.* To prove this theorem, we start with a set $J \notin \mathcal{G}_\mu$ and find necessary conditions for such set to belong to $\mathcal{G}_s$. These necessary conditions, expressed as inequalities in terms of smooth and continuous functions of the $\hat{N}_d / N$ and the $\hat{t}_d / N$, are then used to bound the probability of interest.

Finally, we use Theorem 5.4.3 in Fuller (1996) to show that this probability converges to zero with a rate of $O(n^{-1})$.

Let $\mathbf{A}_\mu$, $\mathbf{A}_{\mu,J}$ and $\boldsymbol{\gamma}_{\mu_d}$ be the analogous versions of $\mathbf{A}_s$, $\mathbf{A}_{s,J}$ and $\boldsymbol{\gamma}_{s_d}$ obtained by substituting $\tilde{\mathbf{y}}_s$ and $\mathbf{W}_s$ by $\boldsymbol{\mu}$ and $\mathbf{W}_\mu$, respectively. Lemma 2 ensures that both $\mathbf{A}_s$ and $\mathbf{A}_\mu$ are irreducible since $\mathbf{A}$ is.

First, suppose $\varnothing \notin \mathcal{G}_\mu$ and let $J = \varnothing$. Then, from conditions in (2.8), $\varnothing \in \mathcal{G}_s$ if and only if $\langle \tilde{\mathbf{z}}_s, \boldsymbol{\gamma}_{s_j} \rangle \le 0$ for $j = 1, 2, \ldots, m$. In contrast, suppose that $\langle \mathbf{z}_\mu, \boldsymbol{\gamma}_{\mu_j} \rangle \le 0$ for $j = 1, 2, \ldots, m$. Hence, $\varnothing \in \mathcal{G}_\mu$, which contradicts our choice of $J$. Therefore, there exists $j_0$ such that $\langle \mathbf{z}_\mu, \boldsymbol{\gamma}_{\mu_{j0}} \rangle > 0$. Then, we have

$$\begin{aligned} P(\varnothing \in \mathcal{G}_s) \le P\left(0 \ge \langle \tilde{\mathbf{z}}_s, \boldsymbol{\gamma}_{s_{j0}} \rangle\right) &= P\left(\langle \mathbf{z}_\mu, \boldsymbol{\gamma}_{\mu_{j0}} \rangle - \langle \tilde{\mathbf{z}}_s, \boldsymbol{\gamma}_{s_{j0}} \rangle \ge \langle \mathbf{z}_\mu, \boldsymbol{\gamma}_{\mu_{j0}} \rangle\right) \\ &= P\left(\left[\frac{\langle \mathbf{z}_\mu, \boldsymbol{\gamma}_{\mu_{j0}} \rangle - \langle \tilde{\mathbf{z}}_s, \boldsymbol{\gamma}_{s_{j0}} \rangle}{\langle \mathbf{z}_\mu, \boldsymbol{\gamma}_{\mu_{j0}} \rangle}\right]^2 \ge 1\right) \\ &\le \frac{1}{\langle \mathbf{z}_\mu, \boldsymbol{\gamma}_{\mu_{j0}} \rangle^2} \, \mathbb{E}\left[\left(\langle \tilde{\mathbf{z}}_s, \boldsymbol{\gamma}_{s_{j0}} \rangle - \langle \mathbf{z}_\mu, \boldsymbol{\gamma}_{\mu_{j0}} \rangle\right)^2\right] \end{aligned}$$

where the last inequality is obtained by an application of Markov's inequality (see for example Casella and Berger (2002), Section 3.6.1). We show now that the expected value in the last term is $O(n^{-1})$. Note that the expression inside of the expected value in the above inequality is a function of vector $\hat{\mathbf{x}}_s = \left(N^{-1}\hat{t}_1, \ldots, N^{-1}\hat{t}_D, N^{-1}\hat{N}_1, \ldots, N^{-1}\hat{N}_D\right)^{\mathsf{T}}$. Let $f_1(\cdot)$ be such a function (which does not depend on $N$), and denote $\mathbf{x}_\mu = (r_1\mu_1, \ldots, r_D\mu_D, r_1, \ldots, r_D)$. To apply Theorem 5.4.3 in Fuller (1996) with $\alpha = 1$, $s = 2$ and $a_N = O(n^{-1/2})$, first we need to show that the following conditions are satisfied:

(a) $\mathbb{E}\left[(\hat{\mathbf{x}}_s - \mathbf{x}_\mu)^2\right] = O(n^{-1})$.

(b) $f_1$ is uniformly bounded in a closed and bounded sphere $\mathcal{S}$.

(c) $f_1^{(i_1, i_2)}(\mathbf{x})$ is continuous in $\mathbf{x}$ over $\mathcal{S}$, where

$$f_1^{(i_1, \ldots, i_r)}(\mathbf{x}_0) = \frac{\partial^r}{\partial_{x_{i_1}} \ldots \partial_{x_{i_r}}} f_1(\mathbf{x})|_{\mathbf{x} = \mathbf{x}_0}.$$

(d) $\mathbf{x}_\mu$ is an interior point of $\mathcal{S}$.

(e) There is a finite number $K$ such that

$$|f_1^{(i_1, i_2)}(\mathbf{x})| \le K \quad \text{for all} \quad \mathbf{x} \in \mathcal{S},$$
$$|f_1^{(i_1)}(\mathbf{x}_\mu)| \le K \quad \text{and} \quad |f_1(\mathbf{x}_\mu)| \le K.$$

Condition (a) is directly met by Lemma 4 (iv). In addition, Lemma 4 (i)-(ii) guarantees that there exist a constant $M > 1$ such that $|N^{-1}\hat{t}_d| \le M$ and $M^{-1} \le N^{-1}\hat{N}_d \le M$. Hence, there exists a closed and bounded sphere $\mathcal{S}$ that it is contained within these constant bounds. Moreover, from Assumption A3, we can conclude that $\mathbf{x}_\mu \in \mathcal{S}$, so condition (d) is satisfied. To show that condition (b) is met, note that $f_1$ is a continuous function in $\mathcal{S}$ since both $W_s^{-1/2}$ and $\tilde{y}_{s_d}$ exist for any $\mathbf{x} \in \mathcal{S}$. Therefore, the Extreme Value Theorem (see Theorem 4.15 in Rudin (1976)) ensures that $f_1$ is uniformly bounded in $\mathcal{S}$. Conditions (c) and (e) are satisfied since $f_1$ is a continuous rational function in $\mathcal{S}$, implying that $f_1$ is infinitely differentiable and its derivatives are bounded in $\mathcal{S}$. Finally, all conditions (a)-(e) are fulfilled. Therefore, from Theorem 5.4.3 in Fuller (1996), we can conclude that $\mathbb{E}\left[f_1(\mathbf{x})\right] = O\left(n^{-1}\right)$, since $f_1$ and its first derivative with respect to the $N^{-1}\hat{t}_d$ and $N^{-1}\hat{N}_d$ evaluate to zero at $\mathbf{x}_\mu$.

Now, take any $J \neq \varnothing$ such that $J \notin \mathcal{G}_\mu$, and assume that $J \in \mathcal{G}_s$. Theorem 1 guarantees that we can always choose a subset $J^* \subseteq J$ such that $J^* \in \mathcal{G}_s, V_{s,J^*}$ is linearly independent, and $\mathcal{L}\left(V_{s,J^*}\right) = \mathcal{L}\left(V_{s,J}\right)$. Note that $\Pi\left(\tilde{\mathbf{z}}_s \mid \mathcal{L}\left(V_{s,J^*}\right)\right) = \mathbf{A}_{s,J^*}^\top \left(\mathbf{A}_{s,J^*}\mathbf{A}_{s,J^*}^\top\right)^{-1} \mathbf{A}_{s,J^*}\tilde{\mathbf{z}}_s$. Let $\tilde{\mathbf{b}}_{s,J^*} = \left(\mathbf{A}_{s,J^*}\mathbf{A}_{s,J^*}^\top\right)^{-1}\mathbf{A}_{s,J^*}\tilde{\mathbf{z}}_s$. Hence, from conditions in (2.8), we have that $J \in \mathcal{G}_s$ implies that $\tilde{\mathbf{b}}_{s,J^*} \geq \mathbf{0}$, and $\left\langle \tilde{\mathbf{z}}_s - \mathbf{A}_{s,J^*}^\top \tilde{\mathbf{b}}_{s,J^*}, \boldsymbol{\gamma}_{s_j} \right\rangle \leq 0$ for any $j$. Define $\mathbf{b}_{\mu,J^*} = \left(\mathbf{A}_{\mu,J^*}\mathbf{A}_{\mu,J^*}^\top\right)^{-1}\mathbf{A}_{\mu,J^*}\mathbf{z}_\mu$ and assume that $\mathbf{b}_{\mu,J^*} \geq \mathbf{0}$, and $\left\langle \mathbf{z}_\mu - \mathbf{A}_{\mu,J^*}^\top \mathbf{b}_{\mu,J^*}, \boldsymbol{\gamma}_{\mu_j} \right\rangle \leq 0$ for $j = 1, 2, \ldots, m$. These conditions would imply that $J^* \in \mathcal{G}_\mu$, contradicting the original assumption that $J \notin \mathcal{G}_\mu$, since $\mathcal{L}\left(V_{\mu,J^*}\right) = \mathcal{L}\left(V_{\mu,J}\right)$ from Lemma 3. Therefore, either there is an element of $\mathbf{b}_{\mu,J^*}$ that is strictly negative or there exists $j_0$ such that $\left\langle \mathbf{z}_\mu - \mathbf{A}_{\mu,J^*}^\top \mathbf{b}_{\mu,J^*}, \boldsymbol{\gamma}_{\mu_{j_0}} \right\rangle > 0$. Hence, proving that $P\left(J \in \mathcal{G}_s\right) = O\left(n^{-1}\right)$ in any of these two scenarios will conclude the proof.

Suppose the $j_0^{\text{th}}$ element of $\mathbf{b}_{\mu,J^*}$ is strictly negative. That is, $\mathbf{e}_{j_0}^\top \mathbf{b}_{\mu,J^*} < 0$, where $\mathbf{e}_j$ denotes the indicator vector that is 1 for entry $j$ and 0 otherwise. Then, we have

$$\begin{aligned} P\left(J \in \mathcal{G}_s\right) \leq P\left(\mathbf{e}_{j_0}^\top \tilde{\mathbf{b}}_{s,J^*} \geq 0\right) &= P\left(\mathbf{e}_{j_0}^\top \tilde{\mathbf{b}}_{s,J^*} - \mathbf{e}_{j_0}^\top \mathbf{b}_{\mu,J^*} \geq -\mathbf{e}_{j_0}^\top \mathbf{b}_{\mu,J^*}\right) \\ &\leq \frac{1}{\left(\mathbf{e}_{j_0}^\top \mathbf{b}_{\mu,J^*}\right)^2} \mathbb{E}\left[\left(\mathbf{e}_{j_0}^\top \tilde{\mathbf{b}}_{s,J^*} - \mathbf{e}_{j_0}^\top \mathbf{b}_{\mu,J^*}\right)^2\right]. \end{aligned}$$

Denote $f_2\left(\hat{\mathbf{x}}_s\right)$ to the expression inside the above expected value. An analogous argument to the one used for the function $f_1$ can be applied to the rational continuous function $f_2$ over $\mathcal{S}$, to conclude that $\mathbb{E}\left[f_2\left(\hat{\mathbf{x}}_s\right)\right] = O\left(n^{-1}\right)$. Note that we also used the fact that $\mathbf{A}_{s,J^*}\mathbf{A}_{s,J^*}^\top$ is an invertible matrix for any $\mathbf{x} \in \mathcal{S}$.

Lastly, suppose there exists $j_0$ such that $\kappa_{\mathbf{z}_\mu, j_0} = \left\langle \mathbf{z}_\mu - \mathbf{A}_{\mu,J^*}^\top \mathbf{b}_{\mu,J^*}, \boldsymbol{\gamma}_{\mu_{j_0}} \right\rangle > 0$, and denote $\kappa_{\tilde{\mathbf{z}}_\mu, j_0} = \left\langle \tilde{\mathbf{z}}_s - \mathbf{A}_{s,J^*}^\top \tilde{\mathbf{b}}_{s,J^*}, \boldsymbol{\gamma}_{s_{j_0}} \right\rangle$. Then, we have

$$\begin{aligned} P\left(J \in \tilde{\mathcal{G}}_s\right) \leq P\left(0 \geq \kappa_{\tilde{\mathbf{z}}_s, j_0}\right) &= P\left(\kappa_{\mathbf{z}_\mu, j_0} - \kappa_{\tilde{\mathbf{z}}_s, j_0} \geq \kappa_{\mathbf{z}_\mu, j_0}\right) \\ &\leq \frac{1}{\kappa_{\mathbf{z}_\mu, j_0}^2} \mathbb{E}\left[\left(\kappa_{\tilde{\mathbf{z}}_s, j_0} - \kappa_{\mathbf{z}_\mu, j_0}\right)^2\right]. \end{aligned}$$

Denote $f_3\left(\hat{\mathbf{x}}_s\right)$ to the expression inside the above expected value. An analogous argument to the one used for the functions $f_1$, $f_2$ is applied to conclude that $\mathbb{E}\left[f_3\left(\hat{\mathbf{x}}_s\right)\right] = O\left(n^{-1}\right)$.

*Proof of Theorem 3.* Take any $J \in \mathcal{G}_s$ and any domain $d$. Note that the condition $\mathbf{A}\boldsymbol{\mu} \geq \mathbf{0}$ implies that $\varnothing \in \mathcal{G}_\mu$. Then, we can write $\tilde{\theta}_{s_d} - \bar{y}_{U_d}$ as

$$\tilde{\theta}_{s_d} - \bar{y}_{U_d} = \left(\tilde{y}_{s_d} - \bar{y}_{U_d}\right) 1_{J = \varnothing} + \sum_{J_G \in \mathcal{G}_\mu \setminus \varnothing} \left(\tilde{\theta}_{s_d,\, J_G} - \bar{y}_{U_d}\right) 1_{J_G = J} + \sum_{J_G \in \mathcal{G}_\mu^c} \left(\tilde{\theta}_{s_d,\, J_G} - \bar{y}_{U_d}\right) 1_{J_G = J},$$

where we used that $\tilde{\theta}_{s_d,\, \varnothing} = \tilde{y}_{s_d}$. Now, an unfeasible variance estimator $\mathrm{AV}\left(\tilde{\theta}_{s_d,\, J}\right)$ can be written as

$$\mathrm{AV}\left(\tilde{\theta}_{s_d,\, J}\right) = \mathrm{AV}\left(\tilde{y}_{s_d}\right) 1_{J = \varnothing} + \sum_{J_G \in \mathcal{G}_\mu \setminus \varnothing} \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right) 1_{J = J_G} + \sum_{J_G \in \mathcal{G}_\mu^c} \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right) 1_{J = J_G}.$$

Hence,

$$\begin{aligned}
\mathrm{AV}\left(\tilde{\theta}_{s_d,\, J}\right)^{-1/2} \left(\tilde{\theta}_{s_d} - \bar{y}_{U_d}\right) &= \mathrm{AV}\left(\tilde{y}_{s_d}\right)^{-1/2} \left(\tilde{y}_{s_d} - \bar{y}_{U_d}\right) 1_{J = \varnothing} \\
&\quad + \sum_{J_G \in \mathcal{G}_\mu \setminus \varnothing} \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)^{-1/2} \left(\tilde{\theta}_{s_d,\, J_G} - \bar{y}_{U_d}\right) 1_{J = J_G} \\
&\quad + \sum_{J_G \in \mathcal{G}_\mu^c} \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)^{-1/2} \left(\tilde{\theta}_{s_d,\, J_G} - \bar{y}_{U_d}\right) 1_{J = J_G} \\
&= \Bigg[ \mathrm{AV}\left(\tilde{y}_{s_d}\right)^{-1/2} \left(\tilde{y}_{s} - \bar{y}_{U_d}\right) 1_{J = \varnothing} \\
&\qquad + \sum_{J_G \in \mathcal{G}_\mu \setminus \varnothing} \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)^{-1/2} \left(\tilde{\theta}_{s_d,\, J_G} - \theta_{U_d,\, J_G}\right) 1_{J = J_G} \\
&\qquad + \sum_{J_G \in \mathcal{G}_\mu^c} \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)^{-1/2} \left(\tilde{\theta}_{s_d,\, J_G} - \theta_{U_d,\, J_G}\right) 1_{J = J_G} \Bigg] \\
&\quad + \left[ \sum_{J_G \in \mathcal{G}_\mu \setminus \varnothing} \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)^{-1/2} \left(\theta_{U_d,\, J_G} - \bar{y}_{U_d}\right) 1_{J = J_G} \right] \\
&\quad + \left[ \sum_{J_G \in \mathcal{G}_\mu^c} \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)^{-1/2} \left(\theta_{U_d,\, J_G} - \bar{y}_{U_d}\right) 1_{J = J_G} \right] \\
&= c_{1N} + c_{2N} + c_{3N},
\end{aligned}$$

where $\theta_{U_d,\, J_G}$ is the population version of $\tilde{\theta}_{s_d,\, J_G}$. A first order term Taylor expansion of $\tilde{\theta}_{s_d,\, J_G}$ and Assumption A6 allow to conclude that each term of the form

$$\mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)^{-1/2} \left(\tilde{\theta}_{s_d,\, J_G} - \theta_{U_d,\, J_G}\right)$$

converges in distribution to a standard normal distribution. Therefore, $c_{1N}$ also converges to a standard normal distribution. Note that for each $J_G \in \mathcal{G}_\mu^c$,

$$\mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)^{-1/2} \left(\theta_{U_d,\, J_G} - \bar{y}_{U_d}\right) = \left[n \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)\right]^{-1/2} \left[n^{1/2} \left(\theta_{U_d,\, J_G} - \bar{y}_{U_d}\right)\right] = O(n^{1/2}),$$

while $1_{J = J_G} = O_p\left(n^{-1}\right)$ by Theorem 2 (since $J \in \mathcal{G}_s$). Thus, $c_{3N} = O_p\left(n^{-1/2}\right)$. Now, note that $\theta_{U_d,\, J_G} - \bar{y}_{U_d} = O\left(N^{-1/2}\right)$ when $J_G \in \mathcal{G}_\mu \setminus \varnothing$ by Assumption A3. Hence, for any $J_G \in \mathcal{G}_\mu \setminus \varnothing$,

$$\mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)^{-1/2} \left(\theta_{U_d,\, J_G} - \bar{y}_{U_d}\right) = \left[n \mathrm{AV}\left(\tilde{\theta}_{s_d,\, J_G}\right)\right]^{-1/2} \left[n^{1/2} \left(\theta_{U_d,\, J_G} - \bar{y}_{U_d}\right)\right] = O\left(\sqrt{\frac{n}{N}}\right),$$

which implies that $c_{2N} = O\left(\sqrt{\frac{n}{N}}\right)$ (bias term). Thus, by combining these properties of $c_{1N}, c_{2N}$ and $c_{3N}$, we conclude that

$$\mathrm{AV}\left(\tilde{\theta}_{s_d,\,J}\right)^{-1/2}\left(\tilde{\theta}_{s_d} - \bar{y}_{U_d}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(B,\ 1),$$

where $B = O\left(\sqrt{\frac{n}{N}}\right)$.

Now, write the feasible variance estimator $\hat{V}\left(\tilde{\theta}_{s_d,\,J}\right)$ as

$$\hat{V}\left(\tilde{\theta}_{s_d,\,J}\right) = \hat{V}\left(\tilde{y}_{s_d}\right) 1_{J=\varnothing} + \sum_{J_G \in \mathcal{G}_\mu \setminus \varnothing} \hat{V}\left(\tilde{\theta}_{s_d,\,J_G}\right) 1_{J=J_G} + \sum_{J_G \in \mathcal{G}_\mu^c} \hat{V}\left(\tilde{\theta}_{s_d,\,J_G}\right) 1_{J=J_G}.$$

By Assumption A6, we have that $\hat{V}\left(\tilde{\theta}_{s_d,\,J_G}\right) - \mathrm{AV}\left(\tilde{\theta}_{s_d,\,J_G}\right) = o_p\left(n^{-1}\right)$ for any $J_G$, which implies that $\hat{V}\left(\tilde{\theta}_{s_d,\,J}\right)^{1/2} - \mathrm{AV}\left(\tilde{\theta}_{s_d,\,J}\right)^{1/2} = o_p\left(n^{-1/2}\right)$. Hence, an application of Slutsky's theorem allows to replace $\mathrm{AV}\left(\tilde{\theta}_{s_d,\,J}\right)^{-1/2}$ by $\hat{V}\left(\tilde{\theta}_{s_d,\,J}\right)^{-1/2}$.

To prove the last part of this theorem, just note that $\mathbf{A\mu} > \mathbf{0}$ implies $\mathcal{G}_\mu = \{\varnothing\}$. Thus, the term $c_{2N}$ does not exist and the bias term vanishes.

# References

Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics*, 12, 470-482.

Breidt, F.J., Opsomer, J.D. and Sanchez-Borrego, I. (2016). Nonparametric variance estimation under fine stratification: An alternative to collapsed strata. *Journal of the American Statistical Association*, 111(514), 822-833.

Casella, G., and Berger, R.L. (2002). *Statistical Inference*. Duxbury, 2nd Edition.

Fuller, W. (1996). *Introduction to Statistical Time Series*. 2nd Edition, New York: John Wiley & Sons, Inc.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 361-374.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Liao, X., and Meyer, M.C. (2014). `coneproj`: an R package for the primal or dual cone projections with routines for constrained regression. *Journal of Statistical Software*, 61, 1-22.

Meyer, M.C. (1999). An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. *Journal of Statistical Planning and Inference*, 81, 13-31.

Meyer, M.C. (2013). A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics*, 42, 1126-1139.

Oliva-Aviles, C., Meyer, M.C. and Opsomer, J.D. (2019). Checking validity of monotone domain mean estimators. *Canadian Journal of Statistics*, 47(2), 315-331.

Opsomer, J.D., Breidt, F.J., White, M. and Li, Y. (2016). Succesive difference replication variance estimation in two-phase sampling. *Journal of Survey Statistical Methodology*, 4(1), 43-70.

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. New York: John Wiley & Sons, Inc.

Rockafellar, R.T. (1970). *Convex Analysis*. New Jersey: Princeton University Press.

Rudin, W. (1976). *Principles of Mathematical Analysis*. New York: McGraw Hill.

Rueda, C., and Lombardía, M. (2012). Small area semiparametric additive monotone models. *Statistical Modelling*, 12(6), 527-549.

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

Wu, J., Meyer, M.C. and Opsomer, J.D. (2016). Survey estimation of domain means that respect natural orderings. *Canadian Journal of Statistics*, 44(4), 431-444.

# Bayesian hierarchical weighting adjustment and survey inference

**Yajuan Si, Rob Trangucci, Jonah Sol Gabry and Andrew Gelman**[1]

Abstract

We combine weighting and Bayesian prediction in a unified approach to survey inference. The general principles of Bayesian analysis imply that models for survey outcomes should be conditional on all variables that affect the probability of inclusion. We incorporate all the variables that are used in the weighting adjustment under the framework of multilevel regression and poststratification, as a byproduct generating model-based weights after smoothing. We improve small area estimation by dealing with different complex issues caused by real-life applications to obtain robust inference at finer levels for subdomains of interest. We investigate deep interactions and introduce structured prior distributions for smoothing and stability of estimates. The computation is done via Stan and is implemented in the open-source R package `rstanarm` and available for public use. We evaluate the design-based properties of the Bayesian procedure. Simulation studies illustrate how the model-based prediction and weighting inference can outperform classical weighting. We apply the method to the New York Longitudinal Study of Wellbeing. The new approach generates smoothed weights and increases efficiency for robust finite population inference, especially for subsets of the population.

**Key Words:** Weighting; Prediction; Multilevel regression and poststratification; Structured prior.

# 1 Introduction

## 1.1 Background

Design-based and model-based approaches have long been contrasted in survey research (Little, 2004). The former automatically takes into account survey design, while the latter can yield robust inference for small sample estimation. Rao (2011) provides an appraisal of frequentist and Bayesian methods on survey sampling practice. Classical design-based approaches use weights to adjust the sample to the population; see Chen, Elliott, Haziza, Yang, Ghosh, Little, Sedransk and Thompson (2017) for a review of various weighted estimators for a population mean. However, classical survey weighting usually relies on many user-defined choices so that the process of weighting can be difficult to codify in real-world surveys (Gelman, 2007). The Bayesian approach for finite population inference (Ghosh and Meeden, 1997) allows prior information to be incorporated, when appropriate, but is subject to model misspecification.

In the present paper we combine Bayesian prediction and weighting in a unified approach to survey inference, applying scalable and robust Bayesian regression models to account for complex design features under the framework of multilevel regression and poststratification (MRP, Gelman and Little (1997); Park, Gelman and Bafumi (2005); Ghitza and Gelman (2013); Si, Pillai and Gelman (2015)). MRP adjusts for complex design and response mechanisms and improves small area estimation (Fay and Herriot, 1979; Rao and Molina, 2015). We deal with different complex issues caused by real-life applications and much finer levels for subdomain inference of interest. Our method yields efficient and valid finite population inference, especially for subgroups, and constructs model-based weights after smoothing.

[1] Yajuan Si, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor. E-mail: yajuan@umich.edu; Rob Trangucci, Department of Statistics, University of Michigan; Jonah Sol Gabry, Department of Statistics, Columbia University; Andrew Gelman, Departments of Statistics and Political Science, Columbia University.

The contributions of this paper are two folded: 1) as innovative Bayesian methodology developments we develop a new structured prior setting to handle high-order interaction terms; and 2) to improve survey research and operation, we combine Bayesian prediction and weighting as a unified approach to survey inference, accounting for design features in the Bayesian modeling. We generalize MRP for finite population inferences and construct stable and calibrated model-based weights to solve the problems of classical weights. We disseminate the R package `rstanarm` implementing the proposed methods for public use, promoting the model-based approaches in survey research and operational practice. More importantly, the paper builds the groundwork to use MRP in the survey weighting adjustment and data integration, for example, to make inferences with nonprobability surveys. Our proposed methods offer one important and practical tool for designing and weighting survey samples (Valliant, Dever and Kreuter, 2018).

## 1.2 Framework

For a finite population of $N$ units, we denote the variable of interest as $y = (y_1, \ldots, y_N)$ and the inclusion indicator variable as $I = (I_1, \ldots, I_N)$, where $I_i = 1$ if unit $i$ is included in the sample and $I_i = 0$ otherwise. Here, inclusion refers to selection and response. The general inference framework considers the joint distribution for $I$ and $y$. Design-based inference considers the distribution of $I$ and treats $y$ as fixed. Under probability sampling, model-based inferences can be based on the distribution of $y$ alone given the variables that affect the inclusion mechanisms are included in the model (Royall, 1968), that is, under the ignorable inclusion mechanism when the distribution of $I$ given $y$ is independent of the distribution of $y$ (Rubin, 1976, 1983).

To account for the factors that affect inclusion, classical design weights adjust for unequal probabilities of sampling, with subsequent weighting adjustments accounting for coverage problems and nonresponse during data collection or data cleaning. Classical weights are thus generated as a product of multiple adjustment factors: inverse probability of selection, inverse propensity score of response, and poststratification (also called calibration or benchmarking; Holt and Smith (1979)). Each of these adjustments can be approximate when the probability of selection, the probability of response, or population totals are estimated from data. Beyond any approximation issues, even if the inclusion model is known exactly, extreme values of weights will cause high variability and then inferential problems, especially when the weights are weakly correlated with the survey outcome variable (Rao, 1966a, b; Hájek, 1971; Särndal, Swensson and Wretman, 1992). When the weighting process involves poststratification or nonresponse adjustment – where the weights themselves are random variables – the variance estimation will be different from the cases only with fixed design weights. It is nontrivial to analytically derive a variance estimator under the multi-stage weighting adjustment or complex sampling design.

In practice, the construction of survey weights requires somewhat arbitrary decisions of the selection of variables and interactions, pooling of weighting cells, and weight trimming. It can be unclear whether and how to incorporate auxiliary information (Groves and Couper, 1995). Discussion of smoothing and trimming in the survey weighting literature (e.g., Potter, 1988, 1990; Elliott and Little, 2000; Elliott, 2007;

Xia and Elliott, 2016) has focused on estimating the finite population total or mean, with less attention to subdomain estimates. Beaumont (2008) proposes to regress weights on the survey variables and use the predicted values as smoothed weights, where the direction is inspiring but tangential to the inference objective where good inference properties are desired for the survey variable of interest rather than the weights. Borrowing information on survey outcomes potentially increases efficiency and calls for a general framework.

Gelman (2007) recommends regression models including as covariates any variables that affect selection and response, including stratification variables, clusters, and auxiliary information. Any of these approaches can be sensitive to the prior specification for stable estimation; this is the model-based counterpart to the decisions required for smoothing or trimming classical survey weights. Flexible prediction techniques, such as spline functions, penalized regression and tree-based models, have been proposed to accommodate model-assisted survey estimation (Särndal et al., 1992; Wu and Sitter, 2001; Breidt and Opsomer, 2017; McConville and Toth, 2018).

Model-based and model-assisted weighting adjustment methods for finite population total estimation have been compared by Henry and Valliant (2012). The model-based weighting methods in the superpopulation perspective (Valliant, Dorfman and Royall, 2000) use predictions from regression models to derive case weights, where the predictions are based on hierarchical linear regression models with various bias corrections (Chambers, Dorfman and Wehrly, 1993; Firth and Bennett, 1998). Based on the finite population total estimation, model-assisted methods derive case weights mainly from calibration on benchmark variables (Kott, 2009) via the generalized regression estimator (GREG, Deville and Särndal (1992)). However, the case weights derived from regression predictions can be highly variable and even negative and may damage some domain estimates. Model-based approaches play a vital role in small area estimation but are subject to misspecification and need new developments when the number of domains is large and the inclusion mechanism is not simply random.

To protect against model misspecification, Little (1983) recommends modeling differences in the distribution of outcomes across classes defined by differential probabilities of inclusion. Si et al. (2015) construct poststratification cells based on the unique values of inclusion probabilities and build hierarchical models to smooth cell estimates as advocated by Little (1991, 1993).

We propose to use Bayesian hierarchical models accounting for survey design to generate weights that can be used in design-based inference. The inference is well calibrated and valid with good frequentist properties (Little, 2011). For large samples, the inference will parallel with design-based inference. For small samples, the hierarchical model smoothing will stabilize domain estimation and generate robust weighting adjustment.

We use the intrinsic variables that are used for design weight construction, nonresponse adjustment and calibration, assume they are discretized, and construct poststratification cells based on the cross-tabulation. Weights are derived through the regressing survey outcome on variables used for weighting given the poststratification. The inclusion of the outcome variable into weighting and poststratification can avoid model misspecification and potentially increase efficiency (Fuller, 2009). Multilevel model estimates shrink the cell estimates towards the prediction from the regression model. The MRP framework combines multilevel regression and poststratification, accounts for design features in the Bayesian paradigm, and is then well equipped to handle complex design features. Our proposal distinguishes from the model-based weights in the literature by using the poststratification cell structure and improves by smoothing, thus avoiding negative weight values.

Si et al. (2015) incorporate weights into MRP, increasing flexibility and efficiency comparing to the pseudo-likelihood approach (Pfeffermann, 1993). In the present paper we go further, starting from the variables that are used for weighting and constructing model-based weights as byproducts under MRP. We develop a novel prior specification for the regularization to handle potentially large numbers of poststratification cells. The prior setting allows for variable selection and keeps the hierarchical structure among main effects and high-order interaction terms for categorical variables. That is, if one variable is not predictive, then the high-order interactions involved with this variable are also likely to be not predictive, to facilitate model interpretation. McConville and Toth (2018) use tree-based methods to automatically select poststrata based on auxiliary variables that are potentially correlated with the survey outcome. Our proposed structured prior plays a similar role with the recursive partitioning algorithm to facilitate poststrata selection but improves efficiency by partial pooling. We use the smoothed weights and estimates that are more stable than the regression tree estimator, and the Bayesian framework propagates all sources of uncertainty while McConville and Toth (2019) ignore the variance for tree growing and use mean squared error to approximate the variance.

We have implemented the computation in the R package `rstanarm` (Goodrich and Gabry, 2017). The fully Bayesian inference is realized via Stan (Stan Development Team, 2018, 2017), which uses Hamiltonian Monte Carlo sampling with adaptive path lengths (Hoffman and Gelman, 2014). Stan promotes robust model-based approaches by reducing the computational burden of building and testing new models. The `rstanarm` package allows for efficient Bayesian hierarchical modeling and weighting inference. The codes are publicly available and reproducible. Our developed computation software provides the accessible platform and has the potential to support the unified framework for survey inference.

Section 2 introduces the motivating problem of weighting for an ongoing social science survey. We discuss the method in detail Section 3. Section 4 describes the statistical evaluation of model-based prediction and weighting inference, and demonstrate the efficiency gains in comparison with classical weighting. We apply the proposal to the real-life survey in Section 5. Section 6 summarizes the improvement and discusses further extension.

## 2 Motivating application

Our methodological research is motivated by operational weighting practice for ongoing surveys. Our immediate goal is to construct weights for the New York City (NYC) Longitudinal Study of Wellbeing

(LSW; Si and Gelman (2014); Wimer, Garfinkel, Gelblum, Lasala, Phillips, Si, Teitler and Waldfogel (2014)), a survey organized by the Columbia University Population Research Center, aiming to provide assessments of income poverty, material hardship, and child and family wellbeing of city residents.

We use the LSW as an example to illustrate practical weighting issues and our proposed improvement, with the understanding that similar concerns arise in other surveys. The survey includes a phone sample based on random digit dialing and an in-person respondent-driven sample of beneficiaries from Robin Hood philanthropic services and their acquaintances. We focus on the phone survey here as an illustration. The LSW phone survey interviews 2,002 NYC adult residents, including 500 cell phone calls and 1,502 landline telephone calls, where half of the landline samples are from low-income areas defined by zipcode information. The collected baseline samples are followed up every three months. We match the samples to the 2011 American Community Survey (ACS) records for NYC. The discrepancies are mainly caused by the oversampling of the low-income neighborhoods and nonresponse.

The baseline weighting process (Si and Gelman, 2014) adjusts for unequal probability of selection, coverage bias, and nonresponse. Classical weights are products of estimated inverse probability of inclusion and raking ratios (Deville, Särndal and Sautory, 1993). However, practitioners have to make arbitrary or subjective choices on the selection and values of weighting factors. For example, to construct weights for individual adults, we have to weight up respondents from large households, as just one adult per sampled household is included in the sample. Gelman and Little (1998) recommend the square root of the ratio of household sizes to family sizes for this weighting adjustment because using household sizes as weights (for example, ACS Weighting Method, 2014) tend to overcorrect in telephone surveys. The raking operation procedure in practice adjusts for socio-demographic factors without tailoring for particular surveys.

The survey organizers are interested in the aspects of life quality of city residents, such as the percentage of children who live under poverty and material hardship. Thus, it is important to get accurate estimates for subpopulations. We would like to develop an objective procedure and let the collected survey data determine the weighting process. The basic principle is to adjust for all variables that could affect the selection and response into weighting. Ideally, we expect that variables used for weighting should include phone availability (number of landline/cell phones and duration with interrupted service), family structure, household structure, socio-demographics and potentially their high-order interaction terms. However, the ACS records only provide information on family size, age, ethnicity, sex, education and poverty gap (a family poverty measure). Meanwhile, considering the substantive analysis goal, the variables describing the number of elder people in the family, the number of children in the family, and the family size, as well as their interactions with poverty gap are recommended by the survey organizers to be included into the weighting process to balance the distribution discrepancy with the population.

To generate classical weights, we select the raking factors that could affect the selection and response, including sex, age, education, ethnicity, poverty gap, the number of children in the family, the number of elder people in the family, the number of working aged people in the family, the two-way interaction between age and poverty gap, the two-way interaction between the number of persons in the family and poverty gap, the two-way interaction between the number of children in the family and poverty gap, and the two-way interaction between the number of elder people in the family and poverty gap. We collect the marginal distributions from ACS and implement raking adjustment. The generated weights have to be trimmed due to some extreme values.

However, it is possible that the subjective weighting adjustment includes some variables or interactions that are not essentially predictive or does not take account for all the important factors that could be of substantive interest later. The raking adjustment assumes that these factors are independent. This will cause biased domain inference bases on the cross-tabulation if the correlation structure in the sample is different from that in the population. Ideally, we should match based on the joint distribution of these weighting related variables. However, small cell sizes or empty under the deep interactions will lead to extremely large weights that need cell collapsing.

The problems we face with classical weighting for the LSW baseline survey are reflective of problems for most operational weighting practice in real-life surveys, which are often complicated with complex designs, longitudinal structure or multi-stage response mechanisms. The ad-hoc decisions that often go into classical weighting schemes can result in different practitioners generating different sets of weights for the same survey. In order to avoid the need for subjectivity, it is important to propose a model-based weighting procedure that allows the data to select weighting factors. We would like to incorporate these variables used for weighting into the model for survey outcomes for efficiency gains, model their high-order interaction terms under regularized prior setting and generate the weights that can be equally treated as classical weights. A large number of variables used for weighting and deep interactions will cause small weighting cells based on the cross-tabulation. The small weighting cells call for statistical adjustment for smoothness and stability.

MRP have achieved success for domain estimation at much finer levels. Borrowing the strength of hierarchical modeling framework with an informative prior distribution, we should be able to obtain the estimate after smoothing the sparse cells. Poststratification via census information will match the estimate from the sample to the population. The combination of regression and poststratification is similar to the endogenous poststratification concept (Breidt, 2008; Dahlke, Breidt, Opsomer and Keilegom, 2013). We introduce the MRP framework in detail.

# 3 Method

## 3.1 Multilevel regression and poststratification

In the basic setting, we are interested in estimating the population distribution of the survey outcome $y$. When the weighting process is transparent, we can directly include the auxiliary variable $X$ into regression modeling for the survey outcome $y$. Here $X$ is a $q$-dimensional vector of variables that affect the sampling design, nonresponse and coverage. Conditional on $X$, the distribution of inclusion indicator $I$ is ignorable.

The selection of the auxiliary variables and the availability of their joint distributions in the population are the key to success for MRP, and also for all other methods to adjust for the sampling selection and nonresponse bias and yield valid population inferences. We recommend including all variables that potentially affect the sample inclusion, such as design information, paradata, and socio-demographics. One advantage of MRP is to perform variable selection and stabilize weights in contrast to noisy classical weights.

Another practical challenge is that the population distribution of the calibration variables may be unknown. We obtain the joint population control distribution from ACS in our application study. Wang, Rothschild, Goel and Gelman (2015) used the aggregated exit polls, Zhang, Holt, Yun, Lu, Greenlund and Croft (2015) used the census tract-level information and Yougov (Yougov, 2017) used the Current Population Survey to directly obtain such information for the poststratification adjustment. In practice, we recommend to obtain the population information either directly from census or large studies with minimal errors or estimated based on available information in related studies. Some auxiliary variables' population distribution may not be available in the census database, such as the number of phones, and we can estimate from other surveys as reference samples. Reilly, Gelman and Katz (2001) applied models to predict the unknown population poststratification information. When marginal distributions are available, Little and Wu (1991) discuss an equivalent model approach for raking and Si and Zhou (2020) develop a Bayes raking estimation in the population cell size estimation. We discuss extensions to develop an integrative framework accounting for the estimation uncertainty of unknown control information in Section 6. The availability of population control information with high quality and predictive power directly affects inferential validity, either for model-based or design-based approaches.

Under MRP, the auxiliary variables $X$ are discretized, and their cross-tabulation constructs the poststratification cells $j$, with population cell sizes $N_j$ and sample cell sizes $n_j$, for $j = 1, \ldots, J$, where $J$ is the total number of poststratification cells (Little, 1991, 1993; Gelman and Little, 1997; Gelman and Carlin, 2001). Then the total population size is $N = \sum_{j=1}^{J} N_j$, and the sample size is $n = \sum_{j=1}^{J} n_j$.

Poststratification inference is different from design-based inference under stratified sampling by the fact that $n_j$'s are now random functions of the sampling distribution $I$. In the repeated sampling of $I$, there is a nonzero probability that $n_j = 0$ for some $j$. The usual resolution of this problem is to condition on $n_j$'s observed in the realized sample, however, the sample inference is not design-unbiased conditionally on $n_j$'s. The MRP framework assumes a model for $n_j$'s to account for the design feature.

The poststratification implicitly assumes that the units in each cell are included with equal probability. Suppose $\theta$ is the population estimand of interest, such as the overall or domain means, and it can be expressed as a weighted sum over any subset or domain $D$ of the poststrata,

$$\theta = \frac{\sum_{j \in D} N_j \theta_j}{\sum_{j \in D} N_j}, \tag{3.1}$$

where $\theta_j$ is the corresponding estimand in cell $j$. The proposed poststratified estimator will be of the general form,

$$\tilde{\theta}^{\mathrm{PS}} = \frac{\sum_{j \in D} N_j \tilde{\theta}_j}{\sum_{j \in D} N_j}, \tag{3.2}$$

where $\tilde{\theta}_j$ is the corresponding estimate in cell $j$. Various modeling approaches can be used to estimate the cell estimates, such as the flexible nonparametric Bayesian models and machine learning algorithms (Rasmussen and Williams (2006); Hastie, Tibshirani and Friedman (2009)). Here, we illustrate using a hierarchical regression model.

In practice, survey weights are attached to each unit, even though they are not attributes of individual units. It is natural to generate unit-level weights based on the entire survey design, and use the weighted averages of the form, such as $\tilde{\theta} = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} w_i$. Our goal here is to obtain an equivalent set of unit-level weights $w_i$ through a model-based procedure for the estimation of $\tilde{\theta}^{\mathrm{PS}}$ to connect weighting and poststratification. Therefore, regression models can be used to obtain $\tilde{\theta}_j$, poststratification accounts for the population information, and model-based weights are re-derived via the expression (3.2).

In classical regression models, full poststratification is a special case, where the cell estimates are computed separately for each cell without any pooling effect, i.e., no pooling. For example, if we are interested in the population mean, then the cell means will be used as the cell estimates. Generally, classical regression models are conducted on cell characteristics without going to the extreme fitting separately for each cell. If more interactions among the characteristics are included, the resulted weights become more variable. On the other side, complete pooling ignores the heterogeneity among cells. Hierarchical regression models will smooth the variable estimates under partial pooling.

Gelman (2007) uses the exchangeable normal model as an illustration and shows that the poststratification estimate $\tilde{\theta}^{\mathrm{PS}}$ for population mean can be expressed as a weighted average between the cell means and the global mean, which yields the unit weights, also as a weighted average between the completely smoothed weights, $w_j = 1$, and the weights from full poststratification, $w_j = (N_j / N)/(n_j / n)$. Hierarchical poststratification is approximately equivalent to shrinkage of weights through the shrinkage of the parameter estimates. The degree of shrinkage goes to zero as the sample size increases, which implies that estimates from the model are close to the truth under the sampling design. However, further developments are necessary to handle a large number of cells and deep interactions, and rigorously evaluate the performance of model-based weights.

In our application to the LSW study, the variables used for weighting include age (5 categories), ethnicity/race (5 categories), education (4 categories), sex (2 categories), poverty measure (5 categories), family size (4 categories), the number of elder people (3 categories) and the number of children (4 categories), in the family, and this results in $J = 5 \times 5 \times 4 \times 2 \times 5 \times 3 \times 4 \times 4 = 48{,}000$ poststrata. The majority of the poststratification cells will be empty or sparse due to the limited sample size (2,002). The sample cell sizes are unbalanced. Often cells are arbitrarily collapsed or combined (Little, 1993) without theoretical justification. Recent model-based weighting smoothing procedures across cells could not handle such sparse cases (Elliott and Little, 2000). Xia and Elliott (2016) introduced a Laplace prior for weight smoothing across a modest number of poststrata based on inclusion probabilities but ignored the variables used for weighting and their hierarchy structure. Using the MRP framework, we account for the variable hierarchy structure to smooth and pool the estimates across the sparse and unbalanced cell sizes with a novel set of prior distributions.

## 3.2 Structured prior distribution

We introduce a structured prior distribution to improve MRP under the sparse and unbalanced cell structures, thus yielding stable model-based survey weights that account for design information. Suppose the population distribution of $X$ is known, that is, we can obtain $N_j$'s from the external data to describe a joint distribution of the variables used for weighting. Extension to unknown $N_j$'s is discussed in Section 6. In practice, the number of poststratification cells $J$ can be large, even much larger than the sample size $n$. The variables used for weighting could affect the inclusion through a complex relationship or a differential response mechanism. Deep interactions are essential for complex relationship structure, but we cannot include all and have to select the predictive main effects and interactions.

Suppose the collected survey response is continuous, $y_i$, for $i = 1, \ldots, n$, and we are interested in the population mean $\bar{Y}$ estimation. We use $\left(X^{1\top}, \ldots, X^{J\top}\right)^\top$ to represent the $J \times q$ predictor matrix in the population under the poststratification framework. For illustration, assume a normal distribution,

$$y_i \sim N\left(\theta_{j[i]},\ \sigma_y^2\right), \tag{3.3}$$

where $j[i]$ denotes the cell $j$ that unit $i$ belongs to. We can also consider unequal variances, allowing the cell scale $\sigma_y$ to vary across cells, indexed as $\sigma_j$. For the prior specification of $\theta_j$, one choice can be $\theta_j = X^j \beta$, and $\beta$ is assigned with some prior distribution. In the hierarchical regression example of Gelman (2007), a multivariate normal distribution is considered, $y_i \sim N\left(X_i \beta, \Sigma_y\right)$ and $\beta \sim N\left(0, \Sigma_\beta\right)$, where the covariates include all main effects and a few selected two-way interactions in $X$ and the covariance matrix $\Sigma_\beta$ is diagonal with different scales. However, the model is subject to misspecification, and the generated weights could be negative.

Since $X^j$ consists different level indicators of the $q$ discrete auxiliary variables, we can express the population cell mean $\theta_j$ as

$$\theta_j = \alpha_0 + \sum_{k \in S^{(1)}} \alpha_{j,k}^{(1)} + \sum_{k \in S^{(2)}} \alpha_{j,k}^{(2)} + \ldots + \sum_{k \in S^{(q)}} \alpha_{j,k}^{(q)}, \tag{3.4}$$

where $S^{(l)}$ is the set of all possible $l$-way interaction terms, and $\alpha_{j,k}^{(l)}$ represents the $k^{\text{th}}$ of the $l$-way interaction terms in the set $S^{(l)}$ for cell $j$. For example, $\alpha_{j,k}^{(1)}$'s with $k \in S^{(1)}$ refer to the main effects, $\alpha_{j,k}^{(2)}$'s with $k \in S^{(2)}$ being the two-way interaction terms, for cell $j$. This decomposition covers all possible interactions among the $q$ variables. When the cell structure is sparse, variable selection is necessary. In practical applications, we recommend the initial inclusion of covariates and interactions with substantive importance and scientific interest in Model (3.4) and perform Bayesian variable selection under the proposed structured prior setting.

We induce structured prior distributions to be able to handle deep interactions and account for their hierarchy structure, where the high-order interaction terms will be excluded if one of the corresponding main effects is not selected. Larger main effects often lead to larger effects of the involved interaction terms. Ideally, more shrinkage should be put on the high-order interactions than that on the main effects, and the prior setting should reflect the nested structure. The challenge embodies the problem in Bayesian inference for group-level variance parameters in an ANOVA structure (Gelman, 2005, 2006). Volfovsky and Hoff (2014) introduce a class of hierarchical prior distributions for interaction arrays that can adapt to the potential similarity between adjacent levels, where the covariance matrix for the high-order interactions is assumed as a Kronecker product of the covariance matrices of main effects after adjusting relative magnitudes. Our proposal extends by inducing more structure among the variance parameters, more shrinkage and smoothing effect to handle an extremely large number of cells with unbalanced sizes than the generally balanced setting in Volfovsky and Hoff (2014), and improves the computation performance.

We start with independent prior distributions on the regression parameters $\alpha$:

$$\alpha_{j,k}^{(l)} \sim N\left(0,\ \left(\lambda_k^{(l)}\sigma\right)^2\right),$$

where $\lambda_k^{(l)}$ represents the local scale and $\sigma$ is the global error scale, for $k \in S^{(l)}$ and $l = 1, \ldots, q$. The error scale is the same across the main effects and high-order interactions, while the local scales are different. The shrinkage effect is induced through the specification of local scales. We assume the local scale of high-order interactions is the product of those for the corresponding main effects after adjusting relative magnitudes.

$$\lambda_k^{(l)} = \delta^{(l)} \prod_{l_0 \in M^{(k)}} \lambda_{l_0}^{(l)},$$

where $\delta^{(l)}$ is the relative magnitude adjustment and $M^{(k)}$ is the collection of corresponding main effects that construct the $k^{\text{th}}$ $l$-way interaction in the set $S^{(l)}$. For example, the local scale of the three-way interaction among age, sex, and education, middle-aged men with college education, will be the product of those for the main effects on age, sex, and education, that is, the product of the local scale parameters for middle-aged, men, and college educated, respectively.

We use the following hyperpriors on the scale parameters:

error scale: $\sigma \sim$ Cauchy$_+$ (0, 1)

local scale for main effects: $\lambda_k^{(1)} \sim N_+(0, 1)$

local scale for high-order interactions: $\lambda_k^{(l)} = \delta^{(l)} \prod_{l_0 \in M^{(k)}} \lambda_{l_0}^{(1)}$ (3.5)

relative magnitude for high-order interactions: $\delta^{(l)} \sim N_+(0, 1)$, for $l = 2, \ldots, q$.

Here Cauchy₊ and $N_+$ denotes the positive part of the Cauchy and normal distributions, respectively. Gelman (2006) proposes the half-Cauchy prior for the scale parameter in hierarchical models, which has the appealing property that it allows scale values arbitrarily close to 0, with heavy tails allowing large values when supported by the data. When $\lambda_k^{(l)}$ is close to 0, the posterior samples of $\alpha_{j,k}^{(l)}$ are shrunk towards 0. The scale parameter for the high-order interaction terms will be 0 if any of the related scale parameters for the main effects is 0. The overall regularization effect is determined by the error scale and the multiplicative scale parameters of the corresponding main effects. We assign a noninformative prior distribution to the intercept term and weakly informative prior distributions to the two global error scale parameters $(\sigma_y, \sigma)$, where $\sigma_y$ ~ Cauchy₊ (0, 5).

The global-local shrinkage prior can stabilize random effects modeling in small area estimation (Tang, Ghosh, Ha and Sedransk, 2018). Our proposed prior specification features the global-local shrinkage and group selection of all possible level indicators for the same variable, similar to the group lasso (Yuan and Lin, 2006). We achieve the goal of variable selection under the similar specification with the Horseshoe prior distribution (Carvalho, Polson and Scott, 2010) and improve by setting up the group selection and multiplicative scales for high-order interactions for sparsity gains. We introduce weakly informative half-Cauchy prior distributions to error scales and informative half-normal prior distributions to the local scale parameters to improve parameter shrinkage estimation and computation efficiency. When the posterior estimation of the scale parameter is close to 0, indicating the variable is not predictive; post-processing can be done to exclude the variable from poststratification cell construction for dimension reduction. This class of priors allows for variable selection in high dimension and keeps the hierarchical structure among main effects and interactions.

Piironen and Vehtari (2016) recommend the prior choice for the global shrinkage hyperparameters based on prior beliefs about the number of nonzero coefficients in the model. The hierarchy setting with correlated variables requires further investigation. We use the default choice Cauchy₊ (0, 1) and conduct an extensive sensitivity analysis of the hyperparameter specification, where the results do not change.

## 3.3 Model-based weights

We can re-express (3.4) and (3.5) as the exchangeable normal model:

$$\theta_j \sim N(\alpha_0, \sigma_\theta^2), \quad \sigma_\theta^2 = \sum_{l=1}^{q} \sum_{k \in S^{(l)}} (\lambda_k^{(l)} \sigma)^2 . \tag{3.6}$$

Conditional on the variance parameters, the posterior mean in the normal model with normal prior distribution is a linear function of data; thus we can determine *equivalent weights* $w_i^*$'s so that one can re-express the smoothed estimate $\sum_{j=1}^{J} N_j \big/ N\tilde{\theta}_j$ as a classical weighted average, $\sum_{i=1}^{n} w_i^* y_i \big/ \sum_{i=1}^{n} w_i^*$. Combining the posterior mean estimates for $\theta_j$ and the model-based estimate given in Model (3.2), Gelman (2007) derives the equivalent unit weights in cell $j$ that can be used classically.

$$w_j \approx \frac{n_j / \sigma_y^2}{n_j / \sigma_y^2 + 1/\sigma_\theta^2} \cdot \frac{N_j / N}{n_j / n} + \frac{1/\sigma_\theta^2}{n_j / \sigma_y^2 + 1/\sigma_\theta^2} \cdot 1, \tag{3.7}$$

where the model-based weight is a weighted average between full poststratification without pooling (weights of $(N_j / N) / (n_j / n)$) and complete pooling (weights equal to 1). The pooling or shrinkage factor is $1 / \left(1 + n_j \sigma_\theta^2 / \sigma_y^2\right)$, which depends on the group and individual variances as well as sample size in the cell. The model-based weights are random variables, and fully Bayesian inference will propagate the corresponding variability. We collect the posterior mean values and treat as the weights that can be used the same as classical weights.

## 3.4 Computation

The Bayesian hierarchical prediction and weighting inference procedure is reproducible and scalable. We implement the proposed structured prior distributions in the open source R package `rstanarm` (Goodrich and Gabry, 2017). The computation codes are available online (Si, Trangucci and Gabry, 2020) for public use. We present the example code for the real data application in Appendix A to demonstrate the user-friendly and efficient computation interface, where survey practitioners can straightforwardly use and adapt. The fully Bayesian inference is realized via Stan. As open source and user-friendly software, Stan contributes to the wide application of Bayesian modeling. Survey practitioners resist model-based approaches mainly due to computation burden. However, model-based methods are ready to face the new challenges on big survey data, such as unbalanced cell structure, combining multiple surveys and analyzing streaming data. The development of Stan can improve the generalization of the model-based approach and provide the computational platform for the unified survey inference framework.

In our implementation, the Markov chain Monte Carlo samples mix well and the chains converge quickly. The fast computation speed widens the usability of model-based survey inference approaches. The proposed prior specification improves the stability for smoothed weights under partial pooling. We compare the model-based weights with classical weights in Section 4 and 5 to demonstrate the calibration for design-based properties (Little, 2011). Furthermore, we illustrate the proposed improvement for domain estimation under unbalanced and sparse sample cell structure.

# 4 Simulation studies

We evaluate the Bayesian procedure by the design-based properties and demonstrate the validity. We consider two main simulation scenarios: a slightly unbalanced structure with a moderate number of poststratification cells and a very unbalanced structure with a large number of poststratification cells. We evaluate the statistical validity of the model-based and weighted estimation for the finite population and domain inference to demonstrate the improved capability to solve the classical weighting problems. To illustrate the capability of variable selection and hierarchy maintenance and the resulting efficiency gains, we compare the posterior estimation with that under independent prior setting but without the multiplicative scale constraint, which is similar with Horseshoe prior under group specification, called as independent prior distributions in the paper: $\lambda_k^{(l)} \sim N\left(0,\ \left(\sigma_k^{(l)}\right)^2\right)$.

We consider model-based predictions under the structured prior (Str-P) and the independent prior (Ind-P) distributions. For weighted inference, we evaluate the estimation after applying the model-based weights under structured prior (Str-W) setting, model-based weights under independent prior (Ind-W) distributions, weights obtained via raking adjustment (Rake-W), classical poststratification weights (PS-W), and inverse probability of selection weighting (IP-W). We present the graphical diagnosis tools to compare the weights and weighted inference.

We borrow 2011 ACS survey of NYC adult residents treating it as the “population”, and randomly draw samples out of it according to a pre-specified selection model without nonresponse. We collect covariates from ACS and simulate the outcome variable to obtain the true distribution as a benchmark. The details of model specifications for the following scenarios are presented in Appendix B. We implement the raking procedure by balancing the marginal distributions of the calibration variables in the selection model and generate the raking weights. The classical poststratification weights $N_j / n_j$’s are obtained by matching the selected sample cell indices with those of the population cells. The selection model can provide the inverse probability of selection weights by matching the sampled unit indices. We also generate model-based weights under independent prior distributions for the main effects and high-order interaction terms of the ACS variables. The generated weights are normalized to average 1 for comparison convenience.

## 4.1 Slightly unbalanced structure

We first handle slightly unbalanced structure when the number of poststratification cells and the sample cell sizes are moderate. We implement repeated sampling process to investigate the frequentist properties of model-based predictions and weighted inferences. With little shrinkage effect on high-order interactions, the model-based prediction and weighting with structured prior distributions have similar performance with that under independent prior distributions, while outperforming the classical weighting approaches.

Assume three variables are included in the selection and outcome models: age, ethnicity, and education. We discretize the three variables in ACS as *age* (18–34, 35–44, 45–54, 55–64, 65+), *eth* (non-Hispanic white, non-Hispanic black, Asian, Hispanic, other), and *edu* (less than high school, high school, some college, bachelor degree or above). The number of poststratification cells is $5 \times 5 \times 4 = 100$. We assume the outcome depends on deep interactions, including all the main effects, two-way and three-way interaction terms among the three variables; and the selection indicator depends on the three main effects. The specific values of the coefficients are given in Tables B.2-B.3 in Appendix B. The values are set to reflect the strong correlations between the covariate and dependent variables. And the effects are not necessarily similar across the adjacent factor levels, different from the scenarios in Volfovsky and Hoff

(2014). The error scale in the outcome model is set as 1, where the true value is always fully recovered from the posterior estimation. The data generation model is different from the estimation model, but the latter is flexible enough to cover the former since the dependency structure will be recovered by the estimation. The proposal is robust against model misspecification.

We repeat the sampling 500 times. The sample sizes vary between 2,141 and 2,393 with median 2,288. Empty sample cells occur with spread-out selection probabilities (ranging from 0.001 to 0.269) over the repeated sampling process. The number of occupied cells in the sample is between 80 and 93 with median 87. The slightly unbalanced cell structure is common in practical surveys with simple and clean sampling design. The population quantities of interest include the overall mean, domain means across the $13\ (= 5 + 4 + 4)$ marginal levels of three variables and domain mean for nonwhite youths (an example of interaction between age and race/ethnicity). We examine the absolute value of estimation bias, root mean squared error (RMSE), standard error (SE) approximated by the average value of standard deviations (Ave. SD) and nominal coverage rate of the 95% confidence intervals.

The outputs in Figure 4.1 show that the model predictions have the smallest RMSE, the smallest SE with reasonable coverage rates, and comparable bias among all the methods. All variables affecting the outcome and selection mechanism are included in the modeling to satisfy the Bayesian principle for ignorable sampling mechanism. The model will predict all the cell estimates including the empty cells in the sample, fully using the population information and poststratification cell structure. The weighting inference is conditional on the observed units within occupied cells, and thus less efficient than the model predictions. Generally, the model-based weighting inference has smaller RMSE and SE but more reasonable coverage rates than that with classical weighting. Raking adjustment is not valid for the domain estimation with large bias, large RMSE, and poor coverage, even though the selection mechanism depends on only the main effects. The inverse probability of selection weighting inference tends to have large SE but low coverage rates, especially for domain estimation. The poststratification weighting inference is close to the model-based weighting estimation since the domain sizes are modestly large. The cell shrinkage effect towards no weighting is small (between 0 and 0.19 with mean 0.05) under slightly unbalanced design. The number of cases who are less than high school educated is small (around 80), resulting in large estimation bias and SE for the weighting inferences, but not in model-based predictions. The model-based predictions stabilize the small area estimation by smoothing, as shown in Table 4.1 that displays the numerical comparison for the subdomain inference.

Model prediction performs well and similarly under the structured prior distribution or independent prior distribution. This is expected due to the small shrinkage effect. The cell structure is slightly unbalanced, and the outcome and selection models depend on all the main effects and high-order interaction terms. But the structured prior setting yields more efficient inference than the independent prior setting with smaller SE. This improvement is obvious in the very unbalanced design as shown in the following simulation of Section 4.2.

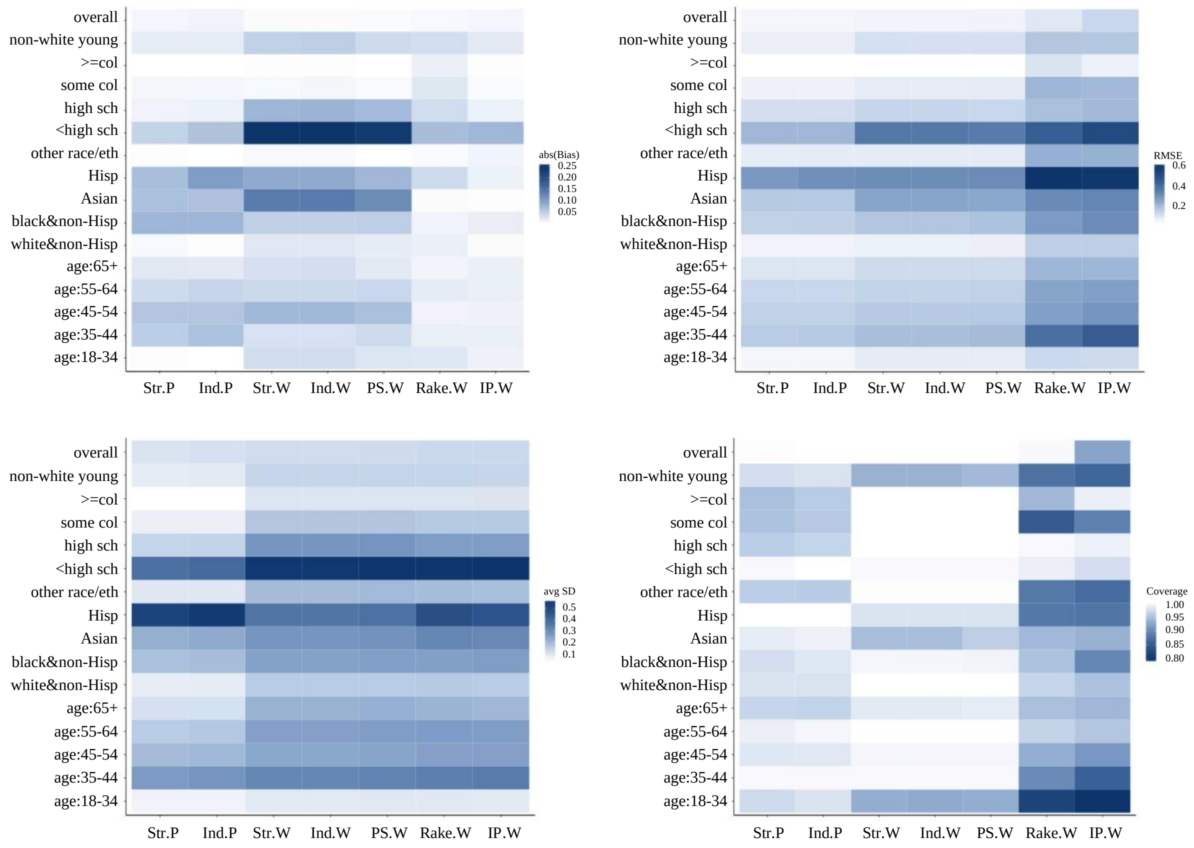


**Figure 4.1 Comparison of prediction and weighting performances on the validity of finite population inference under slightly unbalanced design. The y-axis denotes different groups for the mean estimation. The x-axis includes two model-base prediction methods (Str-P, Ind-P), two model-based weighting methods (Str-W, Ind-W), and three classical weighting methods (PS-W, Rake-W, IP-W). Str-P: model-based prediction under the structured prior; Ind-P: model-based prediction under the independent prior distribution; Str-W: model-based weighting under structured prior; Ind-W: model-based weighting under independent prior distribution; Rake-W: weighting via raking adjustment; PS-W: poststratification weighting; and IP-W: the inverse probability of selection weighting. The plots show that the model-based predictions outperform weighting with the smallest RMSE, the smallest SE, reasonable coverage rates, and comparable bias among all the methods. Model-based weighting inference has smaller RMSE and SE but more reasonable coverage rates than that with classical weighting.**

**Table 4.1**
**Comparison of prediction and weighting performances on the subgroup mean for non-white youth under slightly unbalanced design**

| | Str-P | Ind-P | Str-W | Ind-W | PS-W | Rake-W | IP-W |
|---|---|---|---|---|---|---|---|
| Abs.Bias | 0.02 | 0.02 | 0.05 | 0.05 | 0.04 | 0.03 | 0.02 |
| RMSE | 0.07 | 0.07 | 0.11 | 0.11 | 0.10 | 0.17 | 0.17 |
| Ave.SD | 0.08 | 0.08 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 |
| Coverage | 0.97 | 0.98 | 0.94 | 0.94 | 0.94 | 0.88 | 0.86 |

Additionally, we considered nine cases with different survey outcome models and sample selection models depending on various predictors as in Table B.1 in Appendix B. The specific values of the coefficients are given in Tables B.2-B.3. The conclusions are consistent that the model-based prediction and weighting yield more efficient and precise inference than that under classical weighting, in particular for domain estimation.

## 4.2 Very unbalanced structure

Complex sampling design and response mechanisms tend to create very unbalanced data structures where most poststratification cells are sparse and empty. The proposed structured prior setting brings in strong regularization effect to stabilize the model prediction and improves the estimation efficiency, especially for domain estimation, outperforming the independent prior distributions. The posterior inference on scale parameters can inform variable selection to improve model interpretation. When the main effects are not predictive, neither are the related high-order interactions. However, the posterior inference with independent prior distributions distorts the hierarchical structure between main effects and high-order interactions and hardly informs variable selection. The classical weighting inferences are highly variable in the sparse scenario.

Following the LSW, we collect eight variables in the 2011 ACS-NYC data that affect sample inclusion: *age* (18–34, 35–44, 45–54, 55–64, 65+), *eth* (non-Hispanic white, non-Hispanic black, Asian, Hispanic, other), *edu* (less than high school, high school, some college, bachelor degree or above), *sex* (male, female), *pov* (one household income or poverty measure, poverty gap under 50%, 50–100%, 100–200%, 200–300%, more than 300%), *cld* (0, 1, 2, 3+ young children in the family), *eld* (0, 1, 2+ elders in the family), and *fam* (1, 2, 3, 4+ individuals in the family). The number of unique cells occupied by this classification is 8,874, while the number of poststratification cells constructed by the full cross-tabulation is 48,000.

In the simulation described in Table B.4 and Table B.5, the selection probability depends on the main effects of all variables, while the outcome depends on the main effects of five variables. The cell selection probabilities will be clustered, where some cells have the same selection probabilities. The error scale in the outcome model is set as 1. The selection probabilities fall between 0 and 0.90 with average 0.12, and we select 6,374 units. Even though the sample sizes are large, the simulation creates a very unbalanced structure. The majority of the cells are empty, and 1,096 of 1,925 selected cells have one unit. Starting from an estimation model with sparsity, we assume the Model (3.4) for the cell estimations includes the main effects of the eight variables, eight two-way interactions, and two three-way interactions. These terms are potentially important factors for weighting from the survey organizer’s view. Our proposal can provide the insight of variable selection and then facilitate dimension reduction.

When only the main effects are predictive, the posterior median values under the structured prior setting for the scales of the *cld*, *eld*, and *fam* are small (0.002, 0.003, 0.000), and the posterior median values for the scales of all high-order interactions are close to 0 (with magnitude smaller than or around 0.0001). The posterior mean of the error scale is 0.99 with SE 0.008, close to the true value 1. This is consistent with the simulation design. With independent prior distributions, however, the hierarchical structure between the main effects and high-order interaction terms is ignored. The posterior samples of scale parameters of the high-order interactions can be larger than that of the main effects. It is unclear about their predictive power and then hard to decide which terms to be selected. The posterior samples of the variance parameters under the independent prior distributions tend to be highly variable with heavy tails. For example, the variances of the main effects of age and sex have extremely large sampled values (14,496 and 390,000) and skewed distributions. For variables with a small number of levels, such as sex, the group-level variance estimation is sensitive to the prior distribution, and the independent prior distribution cannot regularize well. The structured prior distribution performs better by assuming the prior distributions share some common parameter and using more information for estimation and then is able to stabilize the variance estimation. The structured prior setting yields more stable inference than the independent prior, and moreover can facilitate variable selection.

The proposed structured prior setting suggests that we exclude the nonpredictive main effects and high-order interactions from the regression model for cell estimates, by either post-processing the posterior samples of the corresponding scales and coefficients to be 0 or refitting the updated model. In the simulation design, three variables affect the selection probability but are not related to the outcome. The inclusion of these variables into the regression model will increase the inference variability. The poststratification cell structure accounts for the eight variables to meet the ignorable sampling assumption. A further modification could be the exclusion of the three variables from the poststratification, which could make the assumption of ignorable sampling vulnerable but have efficiency gains. This is a tradeoff between efficiency and robustness that needs balance based on substantive interest. The selection of survey outcome variables in the weighting process needs further investigation, which we will elaborate in Session 6. We compared the inference with that after excluding the nonpredictive terms and obtained similar outputs for the finite population and domain estimation since the parameter estimates are close to 0 for the nonpredictive terms. Here we present the outputs keeping such variables in the poststratification cell construction and the regression model.

First, we compare the generated weights by the model-based and classical methods. We collect the posterior samples of generated weights and present the posterior mean as the model-based weights. The model-based weights have smaller variability and narrower range than the classical weights, as shown in Figure 4.2. The iterative proportional fitting procedure does not converge after the default 10 iterations that need increasing. We examine the distribution of the outcome after accounting for the weights and compare with the population and sample distribution in the right plot of Figure 4.2. The sample distribution differs from the population distribution by underestimating the outcome values. The weighted distribution shifts towards the true population. The outcome distributions after weighting are similar among the model-based and classical methods, and the model-weights generate a smooth distribution of outcomes. This is reasonable as we expect the model-based weights perform similarly with classical weights on point estimation but improve efficiency by reducing the variability. The shrinkage effect under the structured prior distribution is large, between 0.86 and 1.00 with mean 0.90. The very unbalanced cell structure needs a strong smoothing effect across cells. The model-based weights under the structured prior and independent distributions have similar distributions with the poststratification weights, so the latter two sets of weights are omitted in Figure 4.2.

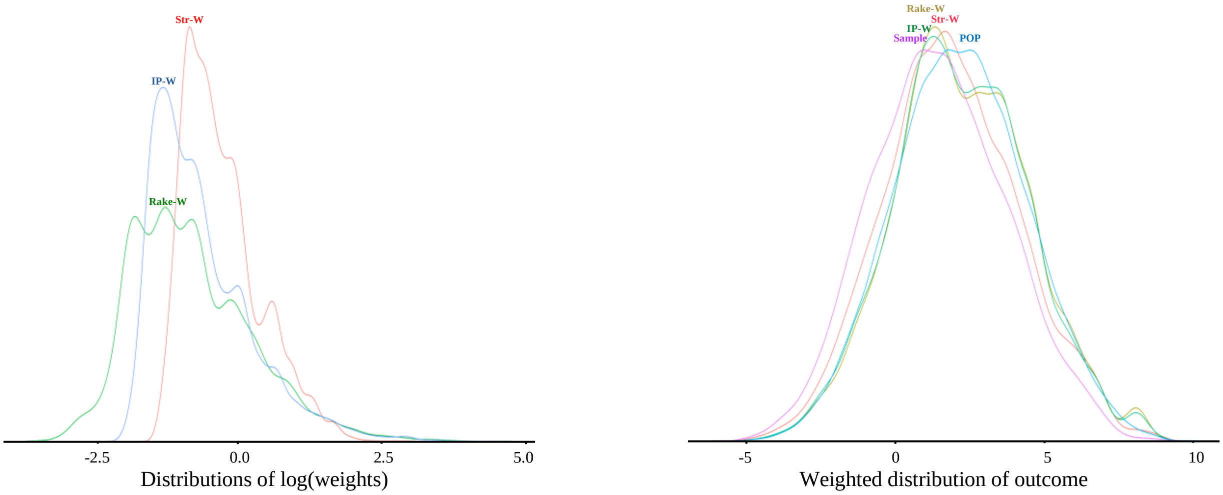


**Figure 4.2 Comparison of generated weights after logarithmic transformation and weighted outcome distributions under a very unbalanced design. Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and IP-W: inverse probability of selection weighting. Sample: sample distribution of the outcome; and POP: population distribution of the outcome. The model-based weights are more stable and generate a more smoothed outcome distribution after weighting than the raking weights and the inverse probability of selection weights.**

We examine the inference for the overall mean and domain means across the marginal levels and for nonwhite young adults. The conclusions are the same as that in Section 4.1. Model-based prediction outperforms weighting inference with smallest bias and SE. The benefit can be explained by that the model uses the population information for empty cell prediction under regularization. Model-based weighting inference has smaller SE than that with classical weighting. Even when the selection probabilities depend on only main effects, raking yields small bias but performs badly with large SE.

Under the very unbalanced design, the model-based weighting inference under structured prior setting is more efficient than that under independent prior setting or with poststratification weights. We compare the SE of the marginal mean estimates of the eight variables from the three weighting methods and plot the relative ratios in the left plot of Figure 4.3. The model-based weighting inference has smaller SE than the poststratification weighting, and the weighting under structured prior setting has the smallest SE. Because the sample sizes and the domain sizes are large and the data generation model is sparse, the model-based weighting inference has a little but not much improvement over the poststratification weighting inference due to small smoothing effect.

The model-based prediction and inference under the structured prior setting are more efficient than that under the independent prior setting. The SEs are smaller with the structured prior than those with the independent prior in the right plot of Figure 4.3. To demonstrate the efficiency gain, we look at the SEs for the population cell estimates. The Bayesian structural inference generally has smaller variability than that with independent prior, especially in the sparse scenarios.

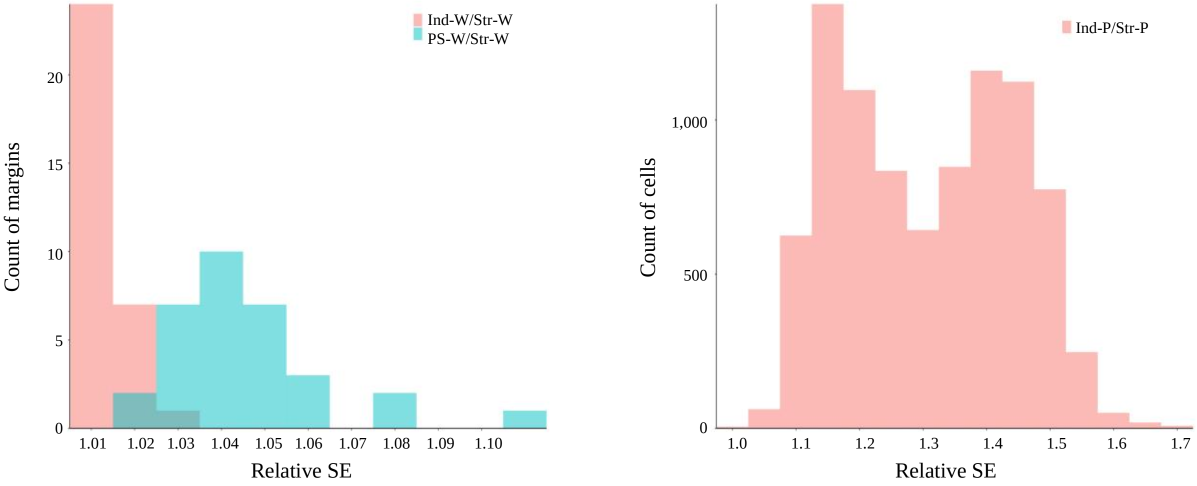


**Figure 4.3 Efficiency comparison of prediction and weighting performances on finite population domain inference under a very unbalanced design. The left plot examines the mean estimation across the margins defined by the eight variables. The right plot presents the population cell mean estimation. The model-based weighting and prediction under the structured prior distribution yield smaller SE than those under independent prior. Model-based weighting yields smaller SE than poststratification weighting.**

We assume different outcome and selection models with different covariates with scenarios summarized in Table B.4 and achieve the same evaluation conclusions.

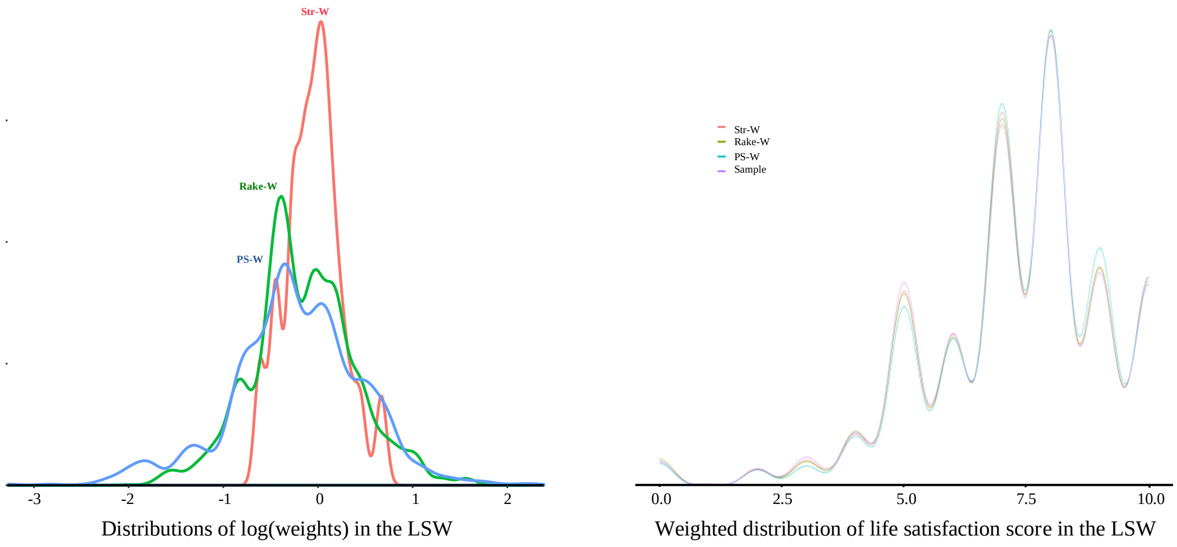
# 5 Application to longitudinal study of wellbeing

With the background introduced in Section 2, we apply the prediction and weighting inference to the NYC Longitudinal Study of Wellbeing. We match the LSW to the adult population via the ACS. We would like to conduct finite population and domain inference and generate weights allowing for general analysis use. The outcome of interest is the self-reported score of life satisfaction on a 1–10 scale. We model the outcome as normally distributed, which is not quite correct given that the responses are discrete, but should be fine in practice for the goal of estimating averages. We first include the same eight variables to construct the poststratification cells and use the same estimation model as those in Section 4.2 under the structured prior setting. The posterior inference shows that the variables *sex*, *cldx*, *eldx*, and *psx* are not predictive, and neither are the related high-order interactions. The scale estimates of such terms have posterior median values close to 0 and several large values as long tails. The posterior samples of scales for several high-order interactions among the remaining four variables concentrate around 0, showing these quantities are not predictive. Another complexity is that, for the sample cells of the LSW, the corresponding population cells are not available in the ACS data. This could happen because the sampling frame is not the ACS survey. The population information is unknown for such cells, and untestable assumptions have to be made. The model fitting improves after variable selection when we check the prediction errors for cell estimates.

Hence, we use four variables after selection, *age*, *eth*, *edu* and *pov*, which constructs 500 poststratification cells. The 2,002 units in the LSW spread out in 359 cells. The largest sample cell has 86 units, while 92 cells have only one unit. The covariates in the model (3.4) for cell estimates include the main effects of the four variables, five two-way interactions (*age* * *eth*, *age* * *edu*, *eth* * *edu*, *age* * *inc* and *eth* * *inc*), and two three-way interactions (*age* * *eth* * *edu* and *age* * *eth* * *inc*). We implement the fully Bayesian inference with the structured prior distributions. We are interested in estimating the average score of life satisfaction for overall and several subgroups of NYC adults and construct weights for general analysis purposes using the LSW.

The posterior median of the unit scale inside cells $\sigma_y$ is 1.93 with 95% credible interval [1.87, 1.99]. The posterior median of the group scale $\sigma_\theta$ is 0.79 with 95% credible interval [0.65, 1.02]. These lead to moderately large shrinkage effects between 0.11 and 0.90 with mean 0.30 across cells. The moderate shrinkage effect makes sense based on the four variables and up to three-way interactions being included. The posterior mean values of the model-based weights are presented in the left plot of Figure 5.1. We can generate the raking weights after adjustment for the marginal distributions of the four variables and poststratification weights based on the ACS data. The population information is obtained after applying the ACS personal weights.

Comparing with the classical weights, our model-based weights have smaller variability with standard deviation 0.32 and the ratio of the maximum and minimum value 3.87, and these values are much smaller than those for the raking and poststratification weights, as shown in Table 5.1. The right plot in Figure 5.1 shows the distribution of the lift satisfaction score after weighting. The model-based weighted distributions and classically weighted distributions are similar as expected, which is consistent with the results in Section 4.2. The weighting process adjusts for the sample distribution by upweighting the high scores and downweighting the low scores. The LSW oversamples poor residents who tend not be satisfied with life, and the weighting adjustment balances the discrepancy.



Distributions of log(weights) in the LSW

Weighted distribution of life satisfaction score in the LSW

**Figure 5.1 Comparison of generated weights after logarithmic transformation and weighted distributions of life satisfaction score in the LSW. Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; IP-W: inverse probability of selection weighting, and Sample: sample distribution of the outcome. The weighted distributions are similar between model-based weights and classical weights, but model-based weights are more stable than classical weights.**

Table 5.1 and Figure 5.2 present the finite population and domain inference. The average score of life satisfaction for NYC adults is 7.24 with standard error 0.05, predicted by the structural model. The estimate is similar to that under model-based weighting and raking inferences, but lower than the poststratification weighting inference. However, the difference is not significant. For example, the structural model predicts the average score of life satisfaction for middle-aged, college-educated whites with income more than three times the poverty level as 7.40 with standard error 0.10, higher than that under weighting inferences. Nevertheless, the predicted scores for the elder with relatively low income (7.37 with SE 0.15) and low-income black New Yorkers (7.01 with SE 0.18) are lower than those under weighting inferences. The discrepancy could be explained by the nonrepresentativeness of the LSW and the deep interactions included by the model. The subgroup of individuals who are middle-aged, college-educated whites may be undercovered in the LSW − as empty poststratification cells occurring − with overcoverage among elderly poor blacks. Weighting the collected samples cannot infer or extrapolate inference on those who are not present in the survey. Though the differences are not significant, inferences conditioning on the collected samples cannot recover the truth, especially for the empty cell estimates. Figure 5.2 shows the model-based prediction yields a higher score for young, highly educated and Hispanic NYC adults, but a lower score for those with poverty gap $< 50\%$, comparing with the weighted inference.

**Table 5.1**
**Comparison of prediction and weighting performances on estimating various domain averages for life satisfaction in the LSW. Str-P: model-based prediction under the structured prior; Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and PS-W: poststratification weighting**

| | Str-P | Str-W | Rake-W | PS-W |
|---|---|---|---|---|
| SD of weights / mean of weights | | 0.32 | 0.66 | 0.80 |
| Max weight / min weight | | 3.87 | 81.28 | 274.65 |
| Overall average for NYC adults ( $n = 2{,}002$) | | | | |
| Est | 7.24 | 7.23 | 7.24 | 7.30 |
| SE | 0.05 | 0.05 | 0.05 | 0.06 |
| Average for middle-aged, college-educated whites with poverty gap $>$ 300% ( $n = 222$ ) | | | | |
| Est | 7.40 | 7.34 | 7.34 | 7.34 |
| SE | 0.10 | 0.11 | 0.11 | 0.11 |
| Average for elders with poverty gap $<$ 200% ( $n = 154$ ) | | | | |
| Est | 7.37 | 7.52 | 7.49 | 7.53 |
| SE | 0.15 | 0.18 | 0.19 | 0.22 |
| Average for blacks with poverty gap $<$ 50% ( $n = 57$ ) | | | | |
| Est | 7.01 | 7.16 | 7.30 | 7.16 |
| SE | 0.18 | 0.26 | 0.28 | 0.29 |

The SEs are similar for the overall mean estimation between predictions and various weighting inferences because of the large sample size. For domain estimation, the model-based prediction and weighting are more efficient than that with raking and poststratification weighting, and the model-based prediction has the smallest standard error. The efficiency gains of model-based prediction and weighting are further demonstrated by domain mean estimation for life satisfaction scores across the marginal levels of four variables, shown in Figure 5.2. The model-based prediction and weighting particularly improve small domain estimation and increase the efficiency.

Survey practitioners often compare the weighted distribution of socio-demographics with the population distribution to check the weighting. While weighting diagnostics need further research and management, we follow this routine to compare the model-based and classical weights. We calculate the Euclidean distances between the weighted distributions and the population distribution for the main effects and high-order interactions among the four variables in the LSW, shown in Table B.6 in Appendix B. The weighted distributions are generally close to the true distributions. Raking focuses on adjusting for the marginal distributions of calibration variables but distorts the joint distributions, where the dependency structure is determined only by the sample without calibration. The poststratification weighting adjusts for the joint distribution, but empty cells in the sample present from the exact matching. The unbalanced cell structure yields unstable inference. The model-based weighting smooths the poststratification weightings and outperforms raking to match the distributions of three-way and four-way interaction terms. Practitioners often rely upon the marginal distributions to evaluate weighting performances, thus in favor of raking. However, raking yields high variable and potentially biased inferences, shown in the Section 4, even in the cases when raking adjustment is correct. Modification of model-based weighting to satisfy such desire on matching marginal distributions will be a future extension to incorporate constraints.

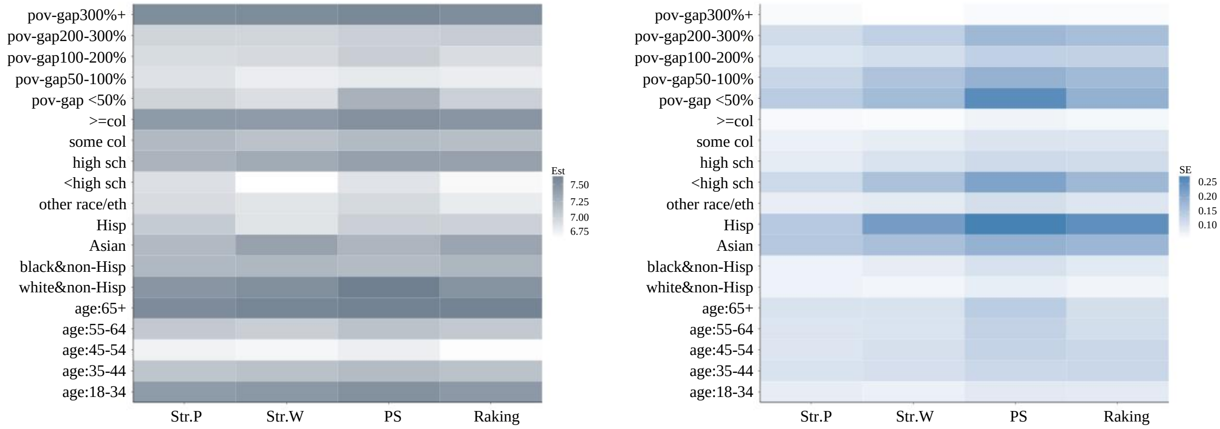


**Figure 5.2 Comparison of predictions and weighting performances on estimating life satisfaction score across the margins of four variables in the LSW. Str-P: model-based prediction under the structured prior; Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and PS-W: poststratification weighting. Model-based predictions and weighting generate different estimates for several subsets and are generally more efficient comparing with classical weighting.**

# 6 Discussion

We combine Bayesian prediction and weighting as a unified approach to survey inference. Multilevel regression with structured prior distributions and poststratification on the population inference yield efficient estimation when accounting for the design feature. The computation is implemented via Stan and disseminated through the R package `rstanarm` for public use, and the software development promotes the model-based approaches in survey research and operational practice. We construct stable and calibrated model-based weights to solve the problems of classical weights. This article builds up the model-based prediction and weighting framework and serves as the first contribution to evaluate the statistical properties of model-based weights and compare the performances with classical weighting. Model-based weights are smoothed across poststratification cells and improve small domain estimation.

The structured prior uses the hierarchical structure between the main effects and high-order interaction terms to introduce multiplicative constraints on the corresponding scale parameters and informs variable selection. Model improvement can be done after post-processing the posterior inferences. The Bayesian structural model yields more stable inference than that with independent prior distributions. Such hierarchy assumption may not be valid for special cases, such as the Exclusive-Or problem where two variables show no main effects but a perfect interaction. However, we do not have strong evidence in the application studies against the plausibility of hierarchy. Furthermore, the unified prediction and weighting approach is well equipped to deal with complex survey designs and big data in surveys, such as streaming data and combining multiple survey studies.

The general MRP framework is open to flexible modeling strategies. In this article, we illustrate by a regression model with all variables of interest and the high-order interactions and incorporate structured prior distributions for regularization. Other approaches, such as nonparametric models and machine learning tools, can be implemented under the MRP framework, being robust against model misspecification. Si et al. (2015) use Gaussian process regression models to borrow information across poststratification cells based on the distances between the inverse inclusion probability weights. Further extensions include applying such flexible approaches to weight smoothing and deriving the model-based weights.

The broad application opportunities come with various challenges that need further investigation. The model-based weights are outcome dependent, which improves the efficiency but potentially reduces the robustness. Survey organizers prefer a set of weights that can be used for general analysis purpose, without being sensitive to outcome selection. We can compare different weights constructed by several important outcomes and conduct sensitivity analysis. When the model-based weights give different inference conclusions, we recommend choosing the set of weights that generate the most reasonable results, with scientific reasoning and be consistent with the population inference.

The weighted marginal distributions of the calibration variables are a bit different from the population inferences, as in Section 5, which does not meet the usual weighting diagnosis standard of survey organizers. The model-weights tend to match the joint distribution to that in the population, but weight smoothing may bring in bias. Tradeoff constraints can be induced to the model to match the marginal distributions.

Another practical challenge is that the population distribution of the calibration variables may be unknown, that is, the population poststratification cell sizes $N_j$'s are unknown. A supplemental model is needed to allow estimation of this information from the sample and integrated with MRP to propagate all sources of uncertainty as an extension, similar to the framework in Si and Zhou (2020) by incorporating known margins. The model-based predictions and weighting inferences need further extensions to handle discrete outcomes, inference on regression coefficients and non-probability or informative sampling designs (Kim and Skinner, 2013). It will be useful to link these ideas on survey inference with the biostatistical and econometrics literature on inverse propensity score and doubly robust weighting (Kang and Schafer, 2007).

## Acknowledgements

We thank the National Science Foundation, National Institutes of Health, Office of Naval Research, Institute of Education Sciences, and Sloan Foundation for grant support.

# Appendix

## A. Example code

Here we present code for the application described in the data. We have written a function `model_based_cell_weights` to calculate the model-based weights from a fitted `rstanarm` model.

```
model_based_cell_weights <- function(object, cell_table) {
  stopifnot(
    is.data.frame(cell_table),
    colnames(cell_table) == c("N", "n")
  )
  draws <- as.matrix(object)
  Sigma <- draws[, grep("^Sigma\\[", colnames(draws)), drop = FALSE]
  sigma_theta_sq <- rowSums(Sigma)
  sigma_y_sq <- draws[, "sigma"]^2
  Ns <- cell_table[["N"]] # population cell counts
  ns <- cell_table[["n"]] # sample cell counts
  J <- nrow(cell_table)
  N <- sum(Ns)
  n <- sum(ns)
  # implementing equation 7 in the paper (although i did some algebra first to
  # simplify the expression a bit)
  Nsy2 <- N * sigma_y_sq
  ww <- matrix(NA, nrow = nrow(draws), ncol = J)
  for (j in 1:J) {
    ww[, j] <-
      (Nsy2 + n * Ns[j] * sigma_theta_sq) / (Nsy2 + N * ns[j] * sigma_theta_sq)
  }
  return(ww)
}
# prepare population data: acs_ad has age, eth, edu and inc
acs_ad %>%
  mutate(
    cell_id = paste0(age, eth, edu, inc)
  ) -> acs_ad
acs_design <- svydesign(id = ~1, weights = ~perwt, data = acs_ad)
agg_pop <-
  svytable( ~ age + eth + edu + inc, acs_design) %>%
  as.data.frame() %>%
  rename(N = Freq) %>%
  mutate(
    cell_id = paste0(age, eth, edu, inc)
  ) %>%
  filter(cell_id %in% acs_ad$cell_id)
# prepare data to pass to rstanarm
# SURVEYdata has 4 variables used for weighting: age, eth, edu and inc; and outcome Y
dat_rstanarm <-
  SURVEYdata %>%
  mutate(
    cell_id = paste0(age, eth, edu, inc)
  )%>%
  group_by(age, eth, edu, inc) %>%
  summarise(
    sd_cell = sd(Y),
    n = n(),
    Y = mean(Y),
    cell_id = first(cell_id)
  ) %>%
  mutate(sd_cell = if_else(is.na(sd_cell), 0, sd_cell)) %>%
  left_join(agg_pop[, c("cell_id", "N")], by = "cell_id")
# Stan fitting under structured prior in rstanarm
fit <-
  stan_glmer(
    formula =
      Y ~ 1 + (1 | age) + (1 | eth) + (1 | edu) + (1 | inc) +
      (1 | age:eth) + (1 | age:edu) + (1 | age:inc) +
      (1 | eth:edu) + (1 | eth:inc) +
      (1 | age:eth:edu) + (1 | age:eth:inc),
    data = dat_rstanarm, iter = 1000, chains = 4, cores = 4,
    prior_covariance =
      rstanarm::mrp_structured(
        cell_size = dat_rstanarm$n,
        cell_sd = dat_rstanarm$sd_cell,
        group_level_scale = 1,
        group_level_df = 1
      ),
    seed = 123,
    prior_aux = cauchy(0, 5),
    prior_intercept = normal(0, 100, autoscale = FALSE),
    adapt_delta = 0.99
  )
# model-based weighting
cell_table <- fit$data[,c("N","n")]
weights <- model_based_cell_weights(fit, cell_table)
weights <- data.frame(w_unit = colMeans(weights),
                 cell_id = fit$data[["cell_id"]],
                 Y = fit$data[["Y"]],
                 n = fit$data[["n"]]) %>%
        mutate(
            w = w_unit / sum(n / sum(n) * w_unit), # model-based weights
            Y_w = Y * w
        )
with(weights, sum(n * Y_w / sum(n)))# mean estimate
```

# B. Simulation designs

Here we present the simulation designs, coefficient values, and comparison on the weighted distributions of socio-demographics as a supplement to Sections 4 and 5.

**Table B.1**
**Covariates in the outcome (O) and selection (S) models for slightly unbalanced design**

| | Case 1 | | Case 2 | | Case 3 | | Case 4 | | Case 5 | | Case 6 | | Case 7 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | O | S | O | S | O | S | O | S | O | S | O | S | O | S |
| age | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| eth | √ | √ | √ | √ | √ | √ | √ | √ | √ | | | √ | | √ |
| edu | √ | √ | √ | √ | √ | | √ | √ | √ | | √ | √ | √ | |
| age*eth | √ | | | √ | √ | √ | | | | √ | | | | √ |
| age*edu | √ | | | √ | √ | | | | | | √ | | √ | |
| eth*edu | √ | | | √ | √ | | | | | | | | | |
| age*eth*edu | √ | | | √ | √ | | | | | | | | | |

**Table B.2**
**Assumed regression coefficients in the outcome model for the simulation using a slightly unbalanced design**

| | All | Main effects | Two variables |
|---|---|---|---|
| age | (0.5, 1.375, 2.25, 3.125, 4) | (0.5, 1.375, 2.25, 3.125, 4) | (0.5, 1.375, 2.25, 3.125, 4) |
| eth | (-2, -1, 0, 1, 2) | (2, -1, 0, 1, 2) | $\vec{0}$ |
| edu | (3, 2, 1, 0) | (3, 2, 1, 0) | (3, 2, 1, 0) |
| age*eth | (4, 2, 1, 1, 3, 3, 2, 1, 1, 1, 2, 3, 2, 2, 1, 4, 4, 3, 2, 3, 2, 4, 1, 4, 1) | $\vec{0}$ | $\vec{0}$ |
| age*edu | (-2, -1, 2, 2, 1, -2, 2, 1, 0, -2, 1, -2, -1, 2, 1, -1, -1, 2, 0, 2) | $\vec{0}$ | (2, 0, -2, -2, 1, 1, -1, -2, -2, -1, -1, 1, 0, -1, -1, 2, 2, 1, -1, 0) |
| eth*edu | (1, -2, 0, -3, -1, 0, -1, -2, 0, -1, -3, -3, 0, -1, -1, 0, 0, -1, 0, -1) | $\vec{0}$ | $\vec{0}$ |
| age*eth*edu | (-1, -0.5, 0.5, -1, -1, -0.5, -1, 0, -1, 0, -1, 0, 1, 1, 0.5, 1, 1, -1, -1, 0, -1, -0.5, -0.5, -1, 1, -1, -0.5, -1, 1, 0, 0.5, 0.5, 1, 0.5, 1, 1, 1, 0.5, 1, 0, 0, -0.5, 0, 1, -1, -1, 0, -1, -1, -1, -0.5, -0.5, 0, 1, -1, 0, 0, -0.5, 1, -0.5, 0.5, -1, 1, 0, 1, 0, -1, 0, -0.5, 1, -0.5, -1, -0.5, 0, 0.5, -0.5, 1, 0.5, -0.5, 0.5, 0, 1, 0, 1, 0.5, 0.5, 0.5, 0, 0, -0.5, 1, -1, 0, 1, 1, 1, 1, -0.5, -1, -1) | $\vec{0}$ | $\vec{0}$ |

**Table B.3**
**Assumed regression coefficients in the selection model for the simulation using a slightly unbalanced design**

| | All | Main effects | Two variables |
|---|---|---|---|
| Intercept | -2 | -2 | -2 |
| age | (-2, -1.75, -1.5, -1.25, -1) | (0, 0.5, 1, 1.5, 2) | (-2, -1.5, -1, -0.5, 0) |
| eth | (-1, -0.25, 0.5, 1.25, 2) | (-2, -1.5, -1, -0.5, 0) | (-1, -0.5, 0, 0.5, 1) |
| edu | (0, 0.67, 1.33, 2) | (0, 1, 2, 3) | $\vec{0}$ |
| age × eth | (1, 1, -1, 1, -1, 1, -1, 0, 0, -1, 0, 0, -1, 1, 0, 0, -1, 1, 1, -1, -1, 0, 1, -1, 1) | $\vec{0}$ | (-1, 1, 1, 1, -1, -1, -1, 0, -1, -1, -1, -1, 1, -1, -1, 0, 1, 1, -1, 1, -1, -1, 1, 0, 0) |
| age × edu | (0, 1, -1, -1, 0, 1, 1, 0, 1, 0, 1, -1, -1, 1, 1, -1, 0, -1, 1, 1) | $\vec{0}$ | $\vec{0}$ |
| eth × edu | (-1, -1, 0, -1, -1, 1, 1, 1, 1, 0, -1, 0, -1, 0, -1, 1, 0, -1, -1, -1) | $\vec{0}$ | $\vec{0}$ |
| age × eth × edu | (0.8, -0.4, 0.6, -0.2, 0.8, 0.2, 0.4, 0.8, 0.4, -0.6, -0.8, -0.4, -0.8, -0.4, 0.4, -1, 0.6, -0.8, -0.6, 0.6, -0.2, 0.2, 0.6, -0.6, 0, 0, -1, -0.2, 0.6, 0.8, -0.4, 0.2, -0.8, 0.4, 0.6, -0.6, 0.8, 0, 0.2, -1, 1, 0.4, 0, 0.8, -0.2, 0, 0, 0.6, -0.8, -0.8, -0.2, 0.4, -1, -0.8, 1, -0.2, 0, 0.8, 0.6, 0.8, -0.2, -0.2, -0.8, 1, 0.8, 0.8, -0.4, -0.8, 0.4, -0.4, 1, -0.6, -1, -0.6, -0.2, 1, 1, -0.2, 1, 0.6, 0.4, 0.8, 0.2, -0.2, -0.6, 0, 0.8, -0.4, 0.4, 0.4, 0.6, -1, -0.8, -0.8, 1, 1, 0.4, 0.6, 0.4, 0.8) | $\vec{0}$ | $\vec{0}$ |

**Table B.4**
**Covariates in the outcome (O) and selection (S) models for a very unbalanced design**

| | Case 1 | | Case 2 | | Case 3 | | Case 4 | |
|---|---|---|---|---|---|---|---|---|
| | O | S | O | S | O | S | O | S |
| age | √ | √ | √ | √ | √ | √ | √ | √ |
| eth | √ | √ | √ | √ | √ | √ | √ | √ |
| edu | √ | √ | √ | √ | √ | √ | √ | √ |
| sex | √ | √ | √ | √ | √ | √ | √ | √ |
| pov | √ | √ | √ | √ | √ | √ | √ | √ |
| cld | | √ | | √ | | √ | √ | √ |
| eld | √ | √ | | √ | √ | √ | √ | √ |
| fam | √ | √ | | √ | √ | √ | √ | √ |
| age*eth | √ | √ | | | √ | | | √ |
| age*edu | √ | √ | | | √ | | | √ |
| eth*edu | √ | √ | | | √ | | | √ |
| eth*pov | √ | √ | | | √ | | | √ |
| age*pov | √ | √ | | | √ | | | √ |
| pov*fam | √ | √ | | | √ | | | √ |
| pov*eld | √ | √ | | | √ | | | √ |
| pov*cld | | √ | | | | | | √ |
| age*eth*edu | √ | √ | | | √ | | | √ |
| age*eth*pov | √ | √ | | | √ | | | √ |

**Table B.5**
**Assumed regression coefficients in the outcome (O) and selection (S) models for a very unbalanced design**

| | O | S |
|---|---|---|
| age | (2, 0, -2, -2, 1) | (0, 0.75, 1.5, 2.25, 3) |
| eth | (1, -1, -2, -2, -1) | (-1, -0.5, 0, 0.5, 1) |
| edu | (-1, 1, 0, -1) | (0, 0.67, 1.33, 2) |
| sex | (-1, 2) | (-1, 0) |
| pov | (2, 1, -1, 0, -1) | (0, 1, 2, 3, 4) |
| cld | $\vec{0}$ | (-1, -0.33, 0.33, 1) |
| eld | $\vec{0}$ | (-2, -1, 0) |
| fam | $\vec{0}$ | (-1, -0.67, -0.33, 0) |

**Table B.6**
**Euclidean distances between the weighted distributions and the population distribution. Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and PS-W: poststratification weighting**

| | Str-W | PS-W | Rake-W |
|---:|:---:|:---:|:---:|
| age | 0.04 | 0.02 | 0.00 |
| eth | 0.08 | 0.06 | 0.00 |
| edu | 0.08 | 0.03 | 0.00 |
| inc | 0.02 | 0.02 | 0.00 |
| age * eth | 0.05 | 0.03 | 0.05 |
| age * edu | 0.05 | 0.02 | 0.05 |
| age * inc | 0.03 | 0.01 | 0.03 |
| eth * edu | 0.06 | 0.04 | 0.05 |
| eth * inc | 0.04 | 0.04 | 0.03 |
| edu * inc | 0.06 | 0.03 | 0.04 |
| age * eth * edu | 0.03 | 0.02 | 0.05 |
| age * eth * inc | 0.03 | 0.02 | 0.04 |
| age * edu * inc | 0.03 | 0.01 | 0.04 |
| eth * edu * inc | 0.04 | 0.02 | 0.04 |
| age * eth * edu * inc | 0.02 | 0.01 | 0.04 |

# References

ACS Weighting Method (2014). *American Community Survey Design and Methodology*, Chapter 11: Weighting and Estimation. United States Census Bureau.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 539-553.

Breidt, F.J. (2008). Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of Statistics*, 36, 403-427.

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

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, 465-480.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88,260-269.

Chen, Q., Elliott, M.R., Haziza, D., Yang, Y., Ghosh, M., Little, R., Sedransk, J. and Thompson, M. (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, 32(2), 227-248.

Dahlke, M., Breidt, F., Opsomer, J. and Keilegom, I.V. (2013). Nonparametric endogenous poststratification in surveys. *Statistica Sinica*, 23, 189-211.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423), 1013-1020.

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Journal of Official Statistics*, 33(1), 23-34.

Elliott, M.R., and Little, R.J. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16(3), 191-209.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a), 269-277.

Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60, 3-21.

Fuller, W. (2009). *Sampling Statistics*. Hoboken: John Wiley & Sons, Inc.

Gelman, A. (2005). Analysis of variance: Why it is more important than ever (with discusion). *Annals of Statistics*, 33(1), 1-53.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3, 515-533.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153-164.

Gelman, A., and Carlin, J.B. (2001). Poststratification and weighting adjustments. In *Survey Nonresponse*, (Eds., R. Groves, D. Dillman, J. Eltinge and R. Little).

Gelman, A., and Little, T.C. (1997). Poststratifcation into many cateogiries using hierarchical logistic regression. *Survey Methodology*, 23, 2, 127-135. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf.

Gelman, A., and Little, T.C. (1998). Improving on probability weighting for household size. *Public Opinion Quarterly*, 62, 398-404.

Ghitza, Y., and Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3), 762-776.

Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman Hall/CRC Press.

Goodrich, B., and Gabry, J.S. (2017). `rstanarm`: Bayesian applied regression modeling via Stan.https://cran.r-project.org/web/packages/rstanarm/.

Groves, R., and Couper, M. (1995). Theoretical motivation for post-survey nonresponse adjustment in household surveys. *Journal of Offcial Statistics*, 11, 93-106.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling" by D. Basu. In *The Foundations of Survey Sampling*, (Eds., V.P. Godambe and D.A. Sprott), 236. Holt, Rinehart and Winston.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2ndEdition. Springer.

Henry, K., and Valliant, R. (2012). Comparing alternative weight adjustment methods. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Hoffman, M.D., and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1351-1381.

Holt, D., and Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142(1), 33-46.

Kang, J.D.Y., and Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523-539.

Kim, J.K., and Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, 100, 385-398.

Kott, P. (2009). Calibration weighting: Combining probability samples and linear prediction models. In *Handbook of Statistics, Sample Surveys: Design, Methods and Application*, (Eds., D. Pfeffermann and C.R. Rao), Volume 29B. Elsevier.

Little, R. (1983). Comment on “An evaluation of model-dependent and probability-sampling inferences in sample surveys”, by M.H. Hansen, W.G. Madow and B.J. Tepping. *Journal of the American Statistical Association*, 78, 797-799.

Little, R. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Little, R. (1993). Post-stratification: A modeler’s perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R. (2004). To model or not to model? Competing modes of inference for finite population sampling inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

Little, R. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26, 162-174.

Little, R., and Wu, M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association*, 86, 87-95.

McConville, K.S., and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2), 389-413.

Park, D.K., Gelman, A. and Bafumi, J. (2005). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In *Public Opinion in State Politics*, (Ed., J.E. Cohen), Standord University Press.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2), 317-337.

Piironen, J., and Vehtari, A. (2016). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. https://arxiv.org/abs/1610.05559.

Potter, F.A. (1988). Survey of procedures to control extreme sample weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 453-458.

Potter, F.A. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.

Rao, J.N.K. (1966a). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A*, 28(1), 47-60.

Rao, J.N.K. (1966b). On the relative effciency of some estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A*, 28(1), 61-70.

Rao, J.N.K. (2011). Impact of frequentist and bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, 26(2), 240-256.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rasmussen, C.E., and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. MITPress, Cambridge, MA.

Reilly, C., Gelman, A. and Katz, J. (2001). Poststratication without population level information on the poststratifying variable, with application to political polling. *Journal of the American Statistical Association*, 96, 1-11.

Royall, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.

Rubin, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581-592.

Rubin, D.B. (1983). Comment on “An evaluation of model-dependent and probability-sampling inferences in sample surveys”, by M.H. Hansen, W.G. Madow and B.J. Tepping. *Journal of the American Statistical Association*, 78, 803-805.

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

Si, Y., and Gelman, A. (2014). Survey weighting for New York Longitudinal Survey on Poverty Measure. Technical report, Columbia University.

Si, Y., and Zhou, P. (2020). Bayes-raking: Bayesian finite population inference with known margins. *Journal of Survey Statistics and Methodology*, smaa008.

Si, Y., Pillai, N.S. and Gelman, A. (2015). Nonparametric Bayesian weighted sampling inference. *Bayesian Analysis*, 10, 605-625.

Si, Y., Trangucci, R. and Gabry, J.S. (2020). Computation codes for manuscript “Bayesian hierarchical weighting adjustment and survey inference”. https://github.com/yajuansisophie/weighting.

Stan Development Team (2017). Stan modeling language user's guide and reference manual.http://mc-stan.org.

Stan Development Team (2018). Stan: A C++ library for probability and sampling. http://mcstan.org.

Tang, X., Ghosh, M., Ha, N.S. and Sedransk, J. (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 0(0),1-14.

Valliant, R., Dever, J.A. and Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*, 2ndEdition. New York: Springer.

Valliant, R., Dorfman, A. and Royall, R. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

Volfovsky, A., and Hoff, P. (2014). Hierarchical array priors for ANOVA decompositions of cross-classified data. *Annals of Applied Statistics*, 8(1), 19-47.

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with nonrepresentative polls. *International Journal of Forecasting*, 31(3), 980-991.

Wimer, C., Garfinkel, I., Gelblum, M., Lasala, N., Phillips, S., Si, Y., Teitler, J. and Waldfogel, J. (2014). Poverty tracker − Monitoring poverty and well-being in NYC. Columbia Population Research Center and Robin Hood Foundation.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185-193.

Xia, X., and Elliott, M.R. (2016). Weight smoothing for generalized linear models using a Laplace prior. *Journal of Official Statistics*, 32(2), 507-539.

Yougov (2017). Introducing the Yougov referendum model. https://yougov.co.uk.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68, 49-67.

Zhang, X., Holt, J.B., Yun, S., Lu, H., Greenlund, K.J. and Croft, J.B. (2015). Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American Journal of Epidemiology*, 182(2), 127-137.

Survey Methodology, December 2020
Vol. 46, No. 2, pp. 215-241
Statistics Canada, Catalogue No. 12-001-X

# Firth's penalized likelihood for proportional hazards regressions for complex surveys

**Pushpal K. Mukhopadhyay**[1]

Abstract

This article proposes a weight scaling method for Firth's penalized likelihood for proportional hazards regression models. The method derives a relationship between the penalized likelihood that uses scaled weights and the penalized likelihood that uses unscaled weights, and it shows that the penalized likelihood that uses scaled weights have some desirable properties. A simulation study indicates that the penalized likelihood using scaled weights produces smaller biases in point estimates and standard errors than the biases produced by the penalized likelihood using unscaled weights. The weighted penalized likelihood is applied to estimate hazard rates for heart attacks by using a public-use data set from the National Health and Epidemiology Followup Study (NHEFS). SAS® statements to estimate hazard rates using data from complex surveys are given in the appendix.

**Key Words:** Monotone likelihood; Delete-one jackknife; Weight scaling.

## 1 Introduction

The Cox proportional hazards regression model (Cox, 1972) is widely used to analyze survival data. It is a semiparametric model that explains the effect of explanatory variables on hazard rates. The model assumes a linear form for the effect of the explanatory variables but allows an unspecified form for the underlying survivor function. The parameters of the model are estimated by maximizing a partial likelihood (Cox, 1972, 1975).

For estimating canonical parameters in the exponential family distributions, Firth (1993) suggested multiplying the likelihood by the Jeffreys prior to obtain a maximum likelihood estimate that is first-order unbiased. The penalized likelihood is of the form

$$L_p(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|I(\boldsymbol{\beta})|^{0.5}$$

where $L(\boldsymbol{\beta})$ is the unpenalized likelihood, $I$ is the information matrix, and $\boldsymbol{\beta}$ is a vector of regression parameters. Firth's penalized likelihood is a very useful technique in practice, not only to reduce bias but also to correct for monotone likelihoods.

Proportional hazards regression models often suffer from monotone likelihoods, in which the likelihood converges to a finite value but at least one parameter diverges (Heinze, 1999). Firth's penalized likelihood is also used to correct monotone likelihoods and to obtain parameter estimates that converge (Heinze, 1999; Heinze and Schemper, 2001; Heinzel, Rüdiger and Schilling, 2002).

Although Firth's penalized likelihood is useful for reducing biases and for obtaining estimates from monotone likelihoods, the penalized likelihood is not studied for complex surveys involving unequal

[1] Pushpal K. Mukhopadhyay, Advanced Analytics Division, SAS Institute Inc., Cary, NC, U.S.A. E-mail: pushpal.mukhopadhyay@sas.com.

weights. It is reasonable to use a weighted likelihood for complex surveys to compensate for unequal weighting (Fuller, 1975; Binder and Patak, 1994). Survey data sets commonly include design weights or analysis weights for which the sum of the weights is an estimator of the population size. However, these unscaled weights will not appropriately scale the information matrix that is used in the penalty term. It is desirable for proportional hazards regression parameters for survey data to have the following two properties:

- *Invariance:* The point estimates and standard errors for the regression parameters should be invariant to the scale of the weights.
- *Closeness:* The Taylor linearized variance for the estimated regression parameters should be close to the delete-one jackknife variance.

In this article, we first show that if the Firth correction is not used, then both the invariance and closeness are satisfied; but if the Firth correction is used with the unscaled weights, then the point estimates and the standard errors are not invariant to the scale of the weights. That is, if the weights are multiplied by a constant and the Firth correction is used, then the point estimates and standard errors will be different. We then propose a commonsense weight scaling method and demonstrate that the Firth correction using the scaled weights has both properties. The only difference between the scaled and unscaled weights is that the sum of the scaled weights is equal to the sample size, but the sum of the unscaled weights is an estimator of the population size.

## 1.1 Example that uses unscaled weights

We used a data set from a study of 65 myeloma patients who were treated with alkylating agents (Lee, Wei and Amato, 1992) to demonstrate the properties of Firth’s penalized likelihood that uses unscaled weights. Survival times in months were recorded for each patient. Patients who were alive after the study period were considered to be censored. The following variables were available for each patient:

- Time: Survival time in months,
- Vstatus: Patient status, zero or one, indicating whether the patient was alive or dead, respectively,
- LogBUN: Log of blood urea nitrogen level,
- HGB: Blood hemoglobin level.

To create a monotone likelihood, we added a new explanatory variable, Contrived, such that its value at all event times is the largest of all values in the risk set (see the example “Firth’s Correction for Monotone Likelihood” in “The PHREG Procedure” in SAS Institute Inc. (2018)). The variable Contrived has the value 1 if the observed survival time is less than or equal to 65; otherwise it has the value 0.

To demonstrate the effect of weights in Firth’s penalized likelihood, we created three weight variables, w1, w3, and w5, with the values of 1, 1,000, and 100,000 for each observation, respectively. Proportional hazards regression parameters are estimated by maximizing a weighted likelihood as described in Section 1.2. Because $w1$ has the value 1 for all observations, using $w1$ in the analysis is equivalent to performing the unweighted analysis.

We fitted the following two proportional hazards models using the PHREG procedure in SAS/STAT® (see "The PHREG Procedure" in SAS Institute Inc. (2018)):

$$\lambda(t, \mathbf{Z}) = \lambda_0(t) \exp(\beta_1 \text{LogBUN} + \beta_2 \text{HGB})$$

$$\lambda(t, \mathbf{Z}) = \lambda_0(t) \exp(\beta_1 \text{LogBUN} + \beta_2 \text{HGB} + \beta_3 \text{Contrived})$$

where $\lambda(t)$ and $\lambda_0(t)$ are the hazard function and the baseline hazard function, respectively. Firth's penalized likelihood is not required in order to fit the first model without Contrived (the likelihood converged in three iteration steps), but the second model containing the variable Contrived does not converge without the Firth penalty in the likelihood. Table 1.1 displays the value of the likelihood and the three regression coefficients for 14 iterations. Although the objective function and the coefficients for LogBun and HGB converge to a finite value after the fourth iteration, the coefficients for Contrived diverges. This is an example of a monotone likelihood for the variable Contrived. Because of this monotonicity, Firth's penalized likelihood must be used to fit the second model containing Contrived.

**Table 1.1**
**Maximum likelihood iteration history showing a monotone likelihood for the variable Contrived**

| Iteration Number | Likelihood Value | LogBUN | HGB | Contrived |
|---|---|---|---|---|
| 1 | -140.693405 | 1.994882 | -0.084319 | 1.466331 |
| 2 | -137.784163 | 1.679468 | -0.109068 | 2.778361 |
| 3 | -136.971190 | 1.714061 | -0.111564 | 3.938095 |
| 4 | -136.707893 | 1.718174 | -0.112273 | 5.003054 |
| 5 | -136.616426 | 1.718755 | -0.112370 | 6.027436 |
| 6 | -136.583520 | 1.718829 | -0.112382 | 7.036445 |
| 7 | -136.571515 | 1.718839 | -0.112384 | 8.039764 |
| 8 | -136.567113 | 1.718841 | -0.112384 | 9.040985 |
| 9 | -136.565495 | 1.718841 | -0.112384 | 10.041434 |
| 10 | -136.564900 | 1.718841 | -0.112384 | 11.041600 |
| 11 | -136.564681 | 1.718841 | -0.112384 | 12.041660 |
| 12 | -136.564601 | 1.718841 | -0.112384 | 13.041683 |
| 13 | -136.564571 | 1.718841 | -0.112384 | 14.041691 |
| 14 | -136.564560 | 1.718841 | -0.112384 | 15.041694 |

If Contrived is not used as an explanatory variable, then all three sets of weights produce the same point estimates and Taylor linearized variance estimates (Table 1.2). The delete-one jackknife variance estimates are also the same for all three sets of weights. Thus, the point estimates and the standard errors are invariant to the scale of the weights when the Firth correction is not used.

**Table 1.2**
**Parameter estimates and standard errors without the Firth correction for all three sets of weights**

| | Estimate | Std. Err. |
|---|---|---|
| **LogBUN** | 1.674 | 0.583 |
| **HGB** | -0.119 | 0.060 |

However, if the unscaled weights are used, then the point estimates for Contrived are not invariant to the scale of the weights. Table 1.3 displays the parameter estimates for three sets of weights when Contrived is used as an explanatory variable (and Firth's penalized likelihood is applied). Because the likelihood is not monotone (Table 1.1) for LogBun and HGB, the point estimates for these two coefficients are not affected by the scale of the weights.

**Table 1.3**
**Parameter estimates with the Firth correction and unscaled weights**

| | **Weight $w1$** | | **Weight $w3$** | | **Weight $w5$** | |
|---|---|---|---|---|---|---|
| | **Estimate** | **Std. Err.** | **Estimate** | **Std. Err.** | **Estimate** | **Std. Err.** |
| LogBUN | 1.722 | 0.584 | 1.719 | 1.85E-2 | 1.719 | 1.85E-3 |
| HGB | -0.112 | 0.061 | -0.112 | 1.93E-3 | -0.112 | 1.93E-4 |
| Contrived | 3.815 | 1.558 | 10.629 | 1.38 | 14.633 | 1.02 |

If Contrived is not used as an explanatory variable, then the ratio of jackknife standard errors to Taylor linearized standard errors is 1.13 and 1.10 for all three sets of weights for the variables LogBUN and HGB, respectively. Thus the ratio of the jackknife variance to the Taylor linearized variance for the unpenalized likelihood is invariant to the scale of weights, and it is reasonable to expect the ratio to be invariant when the penalized likelihood is used.

## 1.2 A brief review of point and variance estimates for regression parameters for finite populations

Before we discuss the weight scaling method, we briefly review point and variance estimates for regression parameters for proportional hazards regression of complex surveys involving unequal weights. Lin and Wei (1989); Binder (1990, 1992); Lin (2000); and Boudreau and Lawless (2006) discussed pseudo-maximum likelihood estimation of proportional hazard regression parameters for survey data. For a more general description for estimating regression parameters for complex surveys, see Kish and Frankel (1974); Godambe and Thompson (1986); Pfeffermann (1993), Korn and Graubard (1999, Chapter 3), Chambers and Skinner (2003, Chapter 2), and Fuller (2009, Section 6.5). Wolter (2007) described different variance estimation techniques for survey data.

Let $\mathcal{U}_N = \{1, 2, \ldots, N\}$ be the set of indices and let $\mathcal{F}_N$ be the set of values for a finite population of size $N$. The survival time of each member of the finite population is assumed to follow its own hazard function, $\lambda_i(t)$, expressed as

$$\lambda_i(t) = \lambda(t; \mathbf{Z}_i(t)) = \lambda_0(t) \exp\left(\mathbf{Z}_i'(t)\ \boldsymbol{\beta}\right)$$

where $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function, $\mathbf{Z}_i(t)$ is a vector of size $P$ of explanatory variables for the $i^{\text{th}}$ unit at time $t$, and $\boldsymbol{\beta}$ is a vector of unknown regression parameters.

The partial likelihood function introduced by Cox (1972, 1975) eliminates the unknown baseline hazard $\lambda_0(t)$ and accounts for censored survival times. If the entire population is observed, then this partial likelihood function can be used to estimate $\boldsymbol{\beta}$. Let $\boldsymbol{\beta}_N$ be the desired estimator.

Assuming a working model with uncorrelated responses, $\boldsymbol{\beta}_N$ is obtained by maximizing the partial log likelihood,

$$l_N(\boldsymbol{\beta}) = \sum_{i \in \mathcal{U}_N} \log \{\mathrm{L}(\boldsymbol{\beta};\, \mathbf{Z}_i(t),\, t_i)\}$$

with respect to $\boldsymbol{\beta}$, where $\mathrm{L}(\boldsymbol{\beta};\, \mathbf{Z}_i(t),\, t_i)$ is Cox's partial likelihood function.

Assume that a probability sample $A_N$ is selected from the finite population $\mathcal{U}_N$. Let $\pi_i$ be the selection probability and $w_i\,(= \pi_i^{-1})$ be the sampling weight for unit $i$. Further assume that explanatory variables $\mathbf{Z}_i(t)$ and survival time $t_i$ are available for every unit in sample $A_N$. A design unbiased estimator for the finite population log likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i \in A_N} \pi_i^{-1} \log \{\mathrm{L}(\boldsymbol{\beta};\, \mathbf{Z}_i(t),\, t_i)\} = \sum_{i \in A_N} w_i \log \{\mathrm{L}(\boldsymbol{\beta};\, \mathbf{Z}_i(t),\, t_i)\}.$$

A sample-based estimator $\hat{\boldsymbol{\beta}}_N$ for the finite population quantity $\boldsymbol{\beta}_N$ can be obtained by maximizing the partial pseudo-log likelihood $l(\boldsymbol{\beta};\, \mathbf{Z}_i(t),\, t_i)$ with respect to $\boldsymbol{\beta}$. The design-based variance for $\hat{\boldsymbol{\beta}}_N$ is obtained by assuming that the set of finite population values $\mathcal{F}_N$ is fixed.

The weighted Breslow likelihood can be expressed as

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\exp\left(\boldsymbol{\beta}' \sum_{\mathcal{D}_k} w_i \mathbf{Z}_i(t)\right)}{\left\{\sum_{\mathcal{R}_k} w_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))\right\}^{\sum_{\mathcal{D}_k} w_i}}$$

where $\mathcal{R}_k$ is the risk set just before the $k^{\text{th}}$ ordered event time $t_{(k)}$, $\mathcal{D}_k$ is the set of individuals who fail at the $t_{(k)}$, and $K$ is the number of distinct event times.

The point estimates for $\boldsymbol{\beta}$ are obtained by maximizing $l(\boldsymbol{\beta}) = \log[L(\boldsymbol{\beta})]$.

Although the weights are sufficient for estimating regression coefficients for the finite population, stratification and clustering information must also be used to estimate sampling variability. In order to estimate sampling variability, you can use either the Taylor series linearization method or a replication method.

### 1.2.1 Analytic variance estimator using the Taylor series linearization method

The Taylor series linearization method uses a sum of squares of the weighted score residuals to estimate the sampling variability.

Define $\bar{\mathbf{Z}}(\boldsymbol{\beta}, t) = \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta},\, t)}{S^{(0)}(\boldsymbol{\beta},\, t)}$, where

$$S^{(0)}(\boldsymbol{\beta}, t) = \sum_{A_N} w_i I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))$$

and

$$\mathbf{S}^{(1)}(\boldsymbol{\beta}, t) = \sum_{A_N} w_i I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))\, \mathbf{Z}_i(t).$$

The score residual for the $i^{\text{th}}$ subject is

$$\begin{aligned}\mathbf{u}_i(\boldsymbol{\beta}) = {} & \Delta_i\left\{\mathbf{Z}_i(t_i) - \overline{\mathbf{Z}}(\boldsymbol{\beta}, t_i)\right\} \\ & - \sum_{j \in A_N}\left[\Delta_j \frac{w_j I(t_i \geq t_j)\exp\left(\boldsymbol{\beta}'\mathbf{Z}_i(t_j)\right)}{S^{(0)}(\boldsymbol{\beta}, t_j)}\left\{\mathbf{Z}_i(t_j) - \overline{\mathbf{Z}}(\boldsymbol{\beta}, t_j)\right\}\right]\end{aligned}$$

where $\Delta_i$ is the event indicator.

Then the Taylor linearized variance estimator is

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\,\mathrm{G}\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$$

where $\mathcal{I}(\hat{\boldsymbol{\beta}})$ is the observed information matrix and the $p \times p$ matrix $\mathbf{G}$ is defined as

$$\mathbf{G} = \sum_{i,\, j \in A_N :\, i < j} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{\hat{\mathbf{u}}_i}{\pi_i} - \frac{\hat{\mathbf{u}}_j}{\pi_j}\right)'\left(\frac{\hat{\mathbf{u}}_i}{\pi_i} - \frac{\hat{\mathbf{u}}_j}{\pi_j}\right)$$

where $\pi_{ij}$ are the joint inclusion probabilities for units $i$ and $j$.

In particular, for stratified cluster designs in which the PSUs are selected by using a simple random sample without replacement, the $p \times p$ matrix $\mathbf{G}$ reduces to

$$\mathbf{G} = \sum_{h=1}^{H}\frac{n_h(1 - f_h)}{n_h - 1}\sum_{i=1}^{n_h}(\mathbf{e}_{hi+} - \bar{\mathbf{e}}_{h\cdot\cdot})'(\mathbf{e}_{hi+} - \bar{\mathbf{e}}_{h\cdot\cdot})$$

where $\mathbf{e}_{hi+}$ is the weighted sum of the score residuals, $\hat{\mathbf{u}}_{hij}$, in stratum $h$ and PSU $i$; $\bar{\mathbf{e}}_{h\cdot\cdot}$ is the mean of $\mathbf{e}_{hi+}$; $n_h$ is the number of PSUs; and $f_h$ is the sampling fraction in stratum $h$.

These estimators are well studied in the sample survey literature. For example, Binder (1992) and Lin (2000) provide conditions under which $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}(\hat{\beta})$ are consistent. Chambless and Boyle (1985) derived the design-based variance and asymptotic normality for discrete proportional hazards models.

### 1.2.2 Replication variance estimator using the delete-one jackknife method

The jackknife method is a commonly used replication variance estimation method for complex surveys. To create replicates, it deletes (assigns a zero weight to) one PSU at a time from the full sample. In each replicate, the sampling weights of the remaining PSUs are modified by the jackknife coefficient $\alpha_r$. The modified weights are called *replicate weights*.

Let PSU $i_r$ in stratum $h_r$ be omitted from the $r^{\text{th}}$ replicate; then the replicate weights and jackknife coefficients are given by

$$w_{hij}^{(r)} = \begin{cases} 0 & i = i_r \text{ and } h = h_r \\ w_{hij}/\alpha_r & i \neq i_r \text{ and } h = h_r \\ w_{hij} & h \neq h_r \end{cases}$$

and $\alpha_r = \frac{n_{h_r} - 1}{n_{h_r}}$, respectively, for all observation units $j$ in stratum $h$ and PSU $i$. The number of PSUs in stratum $h_r$ is $n_{h_r}$.

The jackknife method can be applied to estimate variances for the estimated regression parameters for Cox's model because the model parameters are solutions of a set of estimating equations that are smooth functions of totals (the corresponding score functions are given in Section 2). Properties of jackknife variance estimators for proportional hazard regression models are discussed in Shao and Tu (1995, Section 8.3).

To apply the jackknife method, model parameters are estimated by using the full sample and by using every replicate sample. Let $\hat{\boldsymbol{\beta}}$ be the estimated proportional hazards regression coefficients from the full sample, and let $\hat{\boldsymbol{\beta}}_r$ be the estimated regression coefficients from the $r^{\text{th}}$ replicate. Then the covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated by

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \sum_{r=1}^{R} \alpha_r (\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})'.$$

If the sampling fractions are not ignorable, then the covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated by

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \sum_{r=1}^{R} \alpha_r (1 - f_r)(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})'$$

where $f_r = \frac{n_{h_r}}{N_{h_r}}$ is the sampling fraction in stratum $h_r$.

In practice, both Taylor linearized variance and jackknife variance estimates are used to construct Wald $t$ confidence intervals with $R - H$ degrees of freedom, where $R$ is the number of PSUs (or the number of replicates) and $H$ is the number of strata.

It is straightforward to show that the jackknife variance estimator is algebraically equivalent to the Taylor linearized estimator for design linear estimators. But for design nonlinear estimators, such as the regression coefficients for proportional hazards regression models, the jackknife method tends to produce slightly higher variance estimates than the Taylor linearized method (Fuller, 2009).

Note that if the full sample estimate suffers from a monotone likelihood, then it is very likely that most replicate samples will also suffer from monotone likelihoods. This will results in many "unusable" replicate estimates.

Survey data analysis procedures in SAS/STAT support both Taylor linearized and replication variance estimation methods (Mukhopadhyay, An, Tobias and Watts, 2008).

# 2 Weight scaling

Let $w_i$ be the weight for unit $i$. We propose to use $\tilde{w}_i = \left(\sum_{A_N} 1 \big/ \sum_{A_N} w_i\right) w_i = \left(n \big/ \sum_{A_N} w_i\right) w_i$ as the scaled weight. By construction, the scaled weights are invariant to the scale of the weight. That is, $\tilde{w}_i^* = \left(n \big/ \sum_{A_N} \gamma w_i\right) \gamma w_i = \left(n \big/ \sum_{A_N} w_i\right) w_i = \tilde{w}_i$ for all $\gamma \neq 0$.

Firth's penalized likelihood is given by $L_p(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) |\mathcal{I}(\boldsymbol{\beta})|^{0.5}$, where $L(\boldsymbol{\beta})$ and $\mathcal{I}(\boldsymbol{\beta})$ are the unpenalized likelihood and information matrix, respectively. The penalized log likelihood is

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + 0.5 \log(|\mathcal{I}(\boldsymbol{\beta})|).$$

In particular, when the scaled weights are used, the Breslow unpenalized log partial likelihood (Breslow, 1974) is

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left\{ \boldsymbol{\beta}' \sum_{i \in \mathcal{D}_k} \tilde{w}_i \mathbf{Z}_i(t_k) - \left( \sum_{i \in \mathcal{D}_k} \tilde{w}_i \right) \log \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_k)) \right\}$$

where $w_i$ is the unscaled weight for unit $i$.

Denote

$$\mathbf{S}_k^{(a)}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_k)) [\mathbf{Z}_i(t_k)]^{\otimes a}$$

where $k$ is the $k^{\text{th}}$-ordered event time, $a = 0, 1, 2, [\mathbf{Z}_i(t_k)]^{\otimes 0}$ is 1, $[\mathbf{Z}_i(t_k)]^{\otimes 1}$ is the vector $\mathbf{Z}_i(t_k)$, and $[\mathbf{Z}_i(t_k)]^{\otimes 2}$ is the matrix $[\mathbf{Z}_i(t_k)][\mathbf{Z}_i(t_k)]'$.

Then the score function is given by

$$\begin{aligned} \mathbf{U}(\boldsymbol{\beta}) &\equiv (U(\beta_1), \ldots, U(\beta_p))' \\ &= \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{k=1}^{K} \left\{ \sum_{i \in \mathcal{D}_k} \tilde{w}_i \mathbf{Z}_i(t_k) - \sum_{i \in \mathcal{D}_k} \tilde{w}_i \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{0}(\boldsymbol{\beta})} \right\} \end{aligned}$$

and the Fisher information matrix is given by

$$\begin{aligned} \mathcal{I}(\boldsymbol{\beta}) &= -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \\ &= \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \tilde{w}_i \left\{ \frac{\mathbf{S}_k^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \left[ \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right] \left[ \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right]' \right\}. \end{aligned}$$

Denote

$$\mathbf{Q}_{kp}^{(a)}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_k)) Z_{i,p}(t_k) [\mathbf{Z}_i(t_k)]^{\otimes a}$$

where $a = 0, 1, 2;\ p = 1, \ldots, P;$ and $\mathbf{Z}_i(t) = (Z_{i,1}(t), \ldots, Z_{i,p}(t))$. Then

$$\begin{aligned} \frac{\partial \mathcal{I}(\boldsymbol{\beta})}{\partial \beta_p} = \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \tilde{w}_i & \left\{ \left[ \frac{\mathbf{Q}_{kp}^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kp}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \frac{\mathbf{S}_k^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right] \right. \\ & - \left[ \frac{\mathbf{Q}_{kp}^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kp}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right] \left[ \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right]' \\ & \left. - \left[ \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right] \left[ \frac{\mathbf{Q}_{kp}^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kp}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right]' \right\} \end{aligned}$$

where $p = 1, \ldots, P$.

Point estimates and Taylor linearized standard errors for the penalized likelihood are obtained from the score functions and the Hessian as described in Section 1.2. The jackknife standard errors are obtained by maximizing the penalized likelihood in every replicate sample.

Appendix 1 shows that under certain regularity conditions, the point estimators obtained by maximizing Firth's penalized likelihood are design-consistent.

## 2.1 Penalized likelihoods and the scale of weights

In this section, we derive a relationship between the penalized log likelihood that uses scaled weights and the penalized log likelihood that uses unscaled weights, and we demonstrate that Firth's penalized likelihood using unscaled weights does not have the invariance property.

Let $l(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w})$ be the log likelihood using weights $\tilde{w},$ and let $l(\boldsymbol{\beta}_{w};\, w)$ be the log likelihood using weights $w,$ where $\tilde{w}_i = \alpha w_i$ for all $i$ and $\alpha \neq 0.$ The Breslow log likelihood can be written as

$$\begin{aligned} l(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) &= \sum_{k=1}^{K} \left\{ \boldsymbol{\beta}_{\tilde{w}}' \sum_{i \in \mathcal{D}_k} \tilde{w}_i \mathbf{Z}_i(t_k) - \left( \sum_{i \in \mathcal{D}_k} \tilde{w}_i \right) \log \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp\left( \boldsymbol{\beta}_{\tilde{w}}' \mathbf{Z}_i(t_k) \right) \right\} \\ &= \sum_{k=1}^{K} \left\{ \boldsymbol{\beta}_{\tilde{w}}' \alpha \sum_{i \in \mathcal{D}_k} w_i \mathbf{Z}_i(t_k) - \left( \alpha \sum_{i \in \mathcal{D}_k} w_i \right) \log \sum_{i \in \mathcal{R}_k} \alpha w_i \exp\left( \boldsymbol{\beta}_{\tilde{w}}' \mathbf{Z}_i(t_k) \right) \right\} \\ &= \alpha \sum_{k=1}^{K} \left\{ \boldsymbol{\beta}_{\tilde{w}}' \sum_{i \in \mathcal{D}_k} w_i \mathbf{Z}_i(t_k) - \left( \sum_{i \in \mathcal{D}_k} w_i \right) \log \sum_{i \in \mathcal{R}_k} w_i \exp\left( \boldsymbol{\beta}_{\tilde{w}}' \mathbf{Z}_i(t_k) \right) \right\} \\ &\quad - \sum_{k=1}^{K} \left( \alpha \sum_{i \in \mathcal{D}_k} w_i \right) \log \alpha \\ &= \alpha l(\boldsymbol{\beta}_{\tilde{w}};\, w) - \sum_{k=1}^{K} \left( \alpha \sum_{i \in \mathcal{D}_k} w_i \right) \log \alpha. \end{aligned}$$

Because the second term on the right-hand side does not contain $\boldsymbol{\beta},$ the derivative and the Hessian of the log likelihood are only a multiplier of $\alpha$ and the parameter estimates and standard errors are invariant to the scale of the weights.

However, the following relation shows that the point estimates that are obtained by maximizing the penalized log likelihood are not invariant to the scale of the weights:

$$\begin{aligned} l_p(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) &= l(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) + 0.5 \log \left| I(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) \right| \\ &= \alpha l(\boldsymbol{\beta}_{\tilde{w}};\, w) + 0.5 \log \left| \alpha I(\boldsymbol{\beta}_{\tilde{w}};\, w) \right| - \sum_{k=1}^{K} \left( \alpha \sum_{i \in \mathcal{D}_k} w_i \right) \log \alpha \\ &= \alpha l(\boldsymbol{\beta}_{\tilde{w}};\, w) + 0.5 \left\{ \log \left| I(\boldsymbol{\beta}_{\tilde{w}};\, w) \right| + p \log \alpha \right\} - \sum_{k=1}^{K} \left( \alpha \sum_{i \in \mathcal{D}_k} w_i \right) \log \alpha \\ &= \alpha \left\{ l(\boldsymbol{\beta}_{\tilde{w}};\, w) + 0.5 \log \left| I(\boldsymbol{\beta}_{\tilde{w}};\, w) \right| \right\} - \sum_{k=1}^{K} \left( \alpha \sum_{i \in \mathcal{D}_k} w_i \right) \log \alpha \\ &\quad + 0.5 \left\{ p \log \alpha + (1 - \alpha) \log \left| I(\boldsymbol{\beta}_{\tilde{w}};\, w) \right| \right\} \\ &= \alpha l_p(\boldsymbol{\beta}_{\tilde{w}};\, w) - \sum_{k=1}^{K} \left( \alpha \sum_{i \in \mathcal{D}_k} w_i \right) \log \alpha \\ &\quad + 0.5 \left\{ p \log \alpha + (1 - \alpha) \log \left| I(\boldsymbol{\beta}_{\tilde{w}};\, w) \right| \right\}. \end{aligned}$$

The additional term in the right hand side of the preceding equation involves the regression parameters. Thus the point estimates and the standard errors are not invariant to the scale of the weights.

By construction, point estimates that use the penalized log likelihood and the scaled weights are invariant to the scale of the weights.

## 2.2 Example that uses scaled weights

Consider the myeloma study described in Section 1.1. We refit the same proportional hazards regression model using LogBUN, HGB, and Contrived as explanatory variables, but now we use scaled weights in constructing Firth's penalized likelihood.

Table 2.1 displays point estimates and standard errors from Firth's penalized likelihood using scaled weights and the Taylor linearized variance estimator. These statistics are invariant to the scale of the weights.

**Table 2.1**
**Parameter estimates and their standard errors using the Taylor linearized method with the Firth correction and scaled weights**

| | Weight $w1$ | | Weight $w3$ | | Weight $w5$ | |
|---|---|---|---|---|---|---|
| | **Estimate** | **Std. Err.** | **Estimate** | **Std. Err.** | **Estimate** | **Std. Err.** |
| **LogBUN** | 1.722 | 0.564 | 1.722 | 0.564 | 1.722 | 0.564 |
| **HGB** | -0.112 | 0.064 | -0.112 | 0.064 | -0.112 | 0.064 |
| **Contrived** | 3.815 | 0.458 | 3.815 | 0.458 | 3.815 | 0.458 |

Standard errors using jackknife replicates are also invariant to the scale of the weights. For replicate variance estimation methods, every set of replicate weights must be scaled using the same scaling factor that is used to scale the full sample weights. Table 2.2 displays point estimates and standard errors from Firth's penalized likelihood using scaled weights and the jackknife replicate variance estimator.

**Table 2.2**
**Parameter estimates and their standard errors using jackknife replicates with the Firth correction and scaled weights**

| | Weight $w1$ | | Weight $w3$ | | Weight $w5$ | |
|---|---|---|---|---|---|---|
| | **Estimate** | **Std. Err.** | **Estimate** | **Std. Err.** | **Estimate** | **Std. Err.** |
| **LogBUN** | 1.722 | 0.653 | 1.722 | 0.653 | 1.722 | 0.653 |
| **HGB** | -0.112 | 0.074 | -0.112 | 0.074 | -0.112 | 0.074 |
| **Contrived** | 3.815 | 0.642 | 3.815 | 0.642 | 3.815 | 0.642 |

Estimates from the penalized log likelihood using the scaled weights also have the closeness property. The ratios of jackknife standard errors to Taylor linearized standard errors are 1.16, 1.17, and 1.40 for all three sets of weights for the variables LogBUN, HGB, and Contrived, respectively (Tables 2.1 and 2.2).

# 3 Applications in complex surveys

Data from complex surveys frequently contain unequal weights, strata, and clusters. It is recommended that the weights and other design features be used in the analysis stage. Weighted data provide a better representation of the study population than unweighted data. In this section, we compare the scaled and unscaled weights to estimate proportional hazards regression coefficients through a simulation study and apply the Firth penalized likelihood using the scaled weights to estimate survival times from a data set from NHEFS.

## 3.1 A simulation study

We performed a small simulation study to compare biases in parameter estimates and standard errors for scaled and unscaled weights using Firth’s penalized likelihood. We used two sampling methods to select samples from a fixed finite population: a simple random sample without replacement (SRS) in which each observation unit gets an equal weight; and a probability proportional to size (PPS) without replacement sample in which the sampling weight for an observation unit depends on the value of a size measure associated with the hazard function for the unit. For the purpose of finite population inference, we treat the estimated proportional hazards regression parameters in the finite population as the “true” parameter values. Biases are measured from these true values.

Finite populations of size 10,000 are generated as follows:

- $\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_{10} \sim \text{Bernoulli}(0.75)$,
- $h = \exp(-0.69Z_1 - 0.69Z_2 - \cdots - 0.69Z_{10})$,
- $u \sim \text{uniform}(0, 1)$,
- $t = \log(u)/h$,
- $c \sim \text{Bernoulli}(v)$,
- $m \sim \text{uniform}(10h, 10h + 0.1)$

where $h$ is the hazard function, $t$ is the survival time, $c$ is a censoring indicator, and $m$ is a size measure for each unit. Six finite populations are generated by using different censoring values $(v = 0.1, 0.3, 0.5, 0.7, 0.8, 0.9)$. See Bender, Augustin and Blettner (2005) for methods of generating survival times. Ten regressors $(\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_{10})$ are generated using Bernoulli distributions to create monotone likelihoods, especially when the sample size is small and the censoring rate is high.

Samples are selected from each finite population by using two sampling methods: simple random samples without replacement; and probability proportional to size samples without replacement, where the variable $m$ is used as the size measure. Four sample sizes are used for each sampling method: 50, 100, 500, and 1,000. Sampling weights for all units for SRS depend only on the sample size, but the sampling weight for a unit for PPS depends on both the sample size and the observed value of the variable $m$ for that corresponding unit. To ensure the same distribution of the censored observations in the sampled data as in the population, samples are selected independently from censored and uncensored units in the population.

Finally, the regression parameters from the proportional hazards regression model

$$\lambda(t,\, \mathbf{Z}) = \lambda_{\mathbf{0}}(\mathbf{t})\exp(\beta_{\mathbf{1}}\mathbf{Z}_{\mathbf{1}} + \beta_{\mathbf{2}}\mathbf{Z}_{\mathbf{2}} + \cdots + \beta_{\mathbf{10}}\mathbf{Z}_{\mathbf{10}})$$

are estimated from each sampled data set, where $\lambda_0(t)$ is the baseline hazard, $t$ is the survival time, and $c$ is the censored indicator. The regression parameters are estimated by maximizing the weighted Firth penalized likelihood. Note that the unpenalized likelihood does not converge in most cases because of the monotonicity of the likelihood in the simulated data. When the likelihood is not monotone, we found that the point estimates obtained by using the penalized likelihood are very close to the point estimates obtained by using the unpenalized likelihood. Heinze and Schemper (2001) reported similar findings for unweighted data.

We compare relative biases in point estimates and standard errors using the jackknife method for scaled and unscaled weights. The relative biases (RBs) are defined below (Sitter, 1992).

Let $\hat{\beta}_s$ be the point estimate and $\hat{v}_s$ be the variance estimate for one component of $\boldsymbol{\beta}$ from data set $s$. Define the following:

Relative bias for point estimates, $\hat{\beta}$,

$$\mathrm{RB}(\hat{\beta}) = S^{-1}\sum_{s=1}^{S}\frac{|(\hat{\beta}_s - \beta_T)|}{|\beta_T|}.$$

Relative bias for variance estimates, $\hat{v}$

$$\mathrm{RB}(\hat{v}) = S^{-1}\sum_{s=1}^{S}\frac{|(\hat{v}_s - \mathrm{MSE}_T)|}{\mathrm{MSE}_T}$$

where the true MSE is

$$\mathrm{MSE}_T(\hat{\beta}) = S^{-1}\sum_{s}(\hat{\beta}_s - \beta_T)^2$$

and $\beta_T$ is the "true" parameter value obtained by fitting the proportional hazards regression model using all units in the finite population. The ratio of RBs is defined as the ratio of the RB using the unscaled weights to the RB using the scaled weights.

The median of ratios of RBs over 5,000 repetitions is displayed in this section. We report the median because there are some "bad" samples in which convergences are questionable even with the Firth correction. These "bad" samples produce few estimates with very large biases. Because of these large biases, the mean of the ratio of RBs is a more unstable statistic than the median. Without the "bad" replicates, the mean and medians are very close. We also noticed that the penalized log likelihood using the unscaled weights produces more of these "false" convergences.

Results for all regressors $\mathrm{Z}_1,\, \mathrm{Z}_2, \ldots,\, \mathrm{Z}_{10}$ are similar. For simplicity, we display results for only two regressors, $\mathrm{Z}_3$ and $\mathrm{Z}_8$.

Ratios of RBs in parameter estimates for unscaled and scaled weights for the variables $Z_3$ and $Z_8$ are displayed in Figures 3.1, 3.2, 3.3, and 3.4. For small sample sizes and a large number of censored observations, RBs using scaled weights are much smaller than RBs using unscaled weights. For large sample sizes, RBs from both weights are similar primarily because the Firth option is not necessary, since the convergence is not an issue with large data sets.

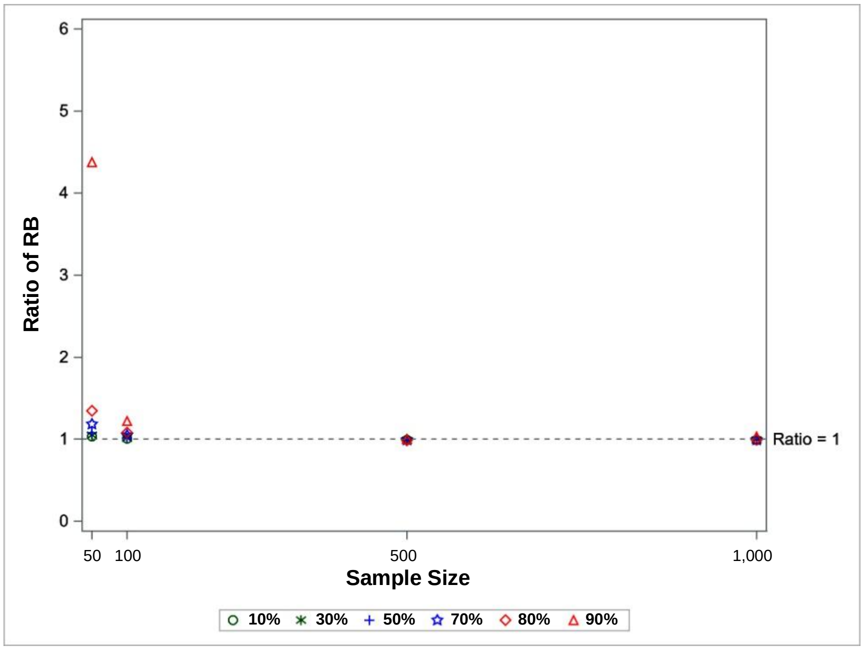


**Figure 3.1 Ratio of relative biases in parameter estimates for SRS samples for Z3.**

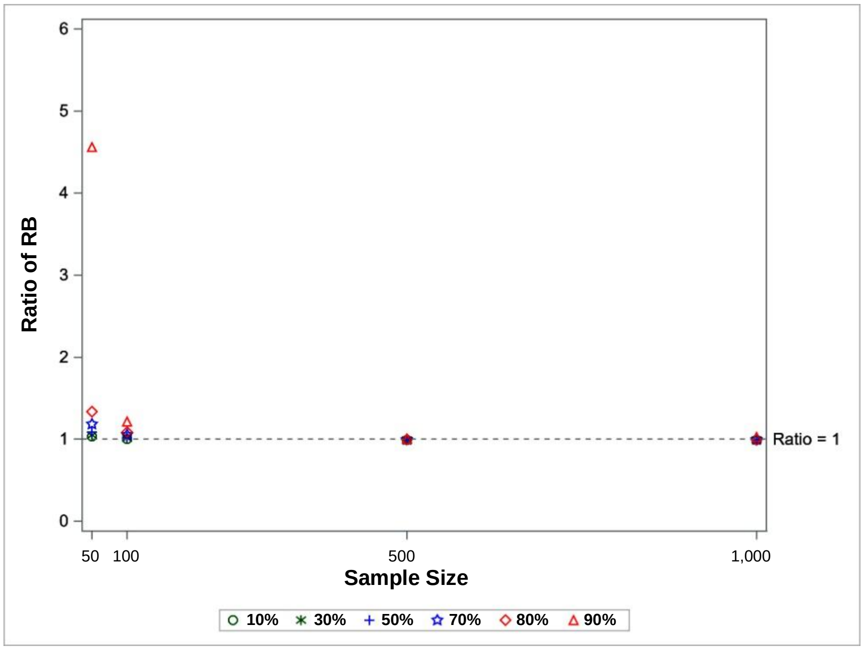


**Figure 3.2 Ratio of relative biases in parameter estimates for SRS samples for Z8.**

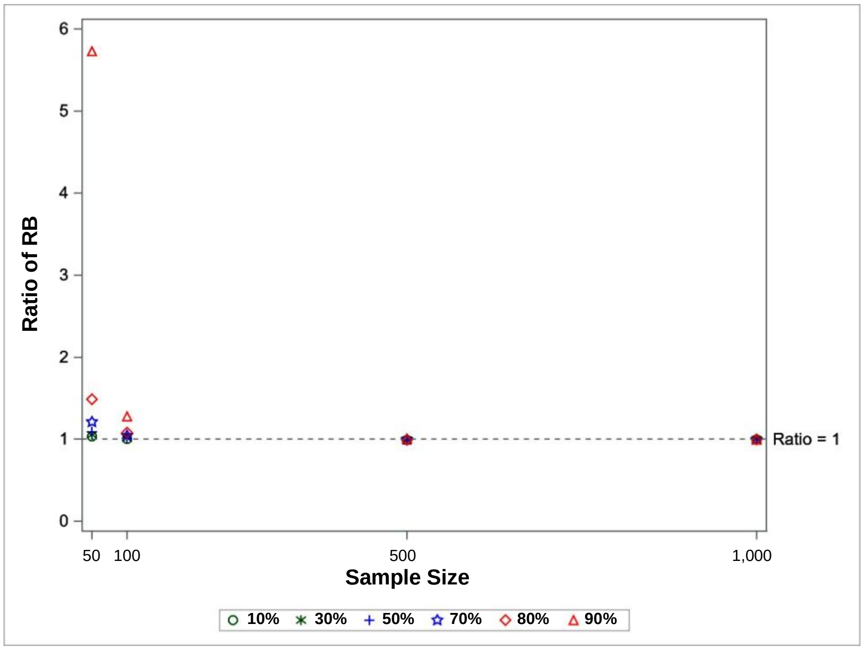


**Figure 3.3 Ratio of relative biases in parameter estimates for PPS samples for Z3.**

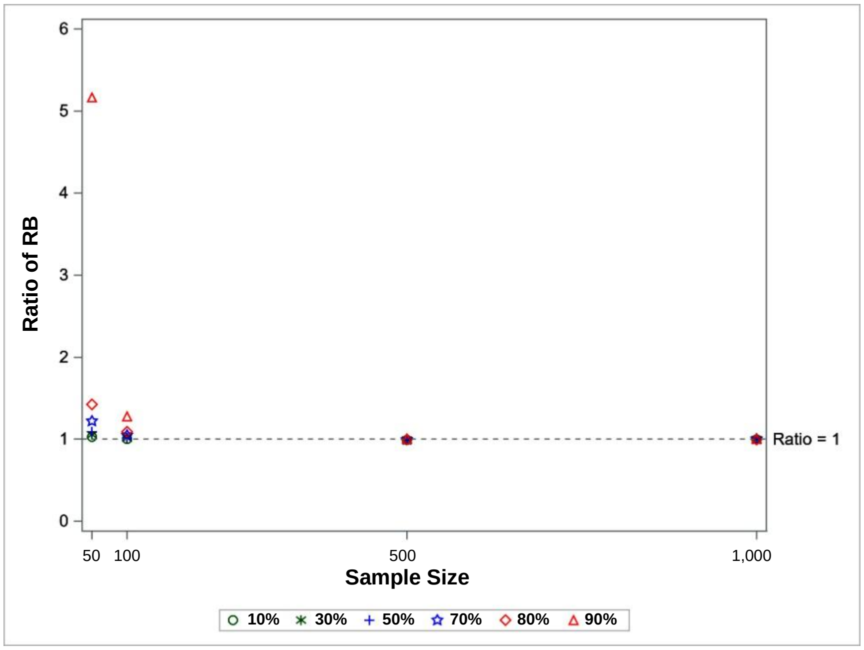


**Figure 3.4 Ratio of relative biases in parameter estimates for PPS samples for Z8.**

Ratios of RBs in standard errors for unscaled and scaled weights for the variables $Z_3$ and $Z_8$ are displayed in Figures 3.5, 3.6, 3.7, and 3.8. RBs for standard errors follow the same trend as RBs for point estimates. However, RBs for standard errors are higher than RBs for point estimates. For small sample sizes and a large number of censored observations, RBs using scaled weights are much smaller than RBs using unscaled weights. For large sample sizes, RBs from both scaled and unscaled weights are similar primarily because the Firth option is not necessary, since the convergence is not an issue with large data sets.

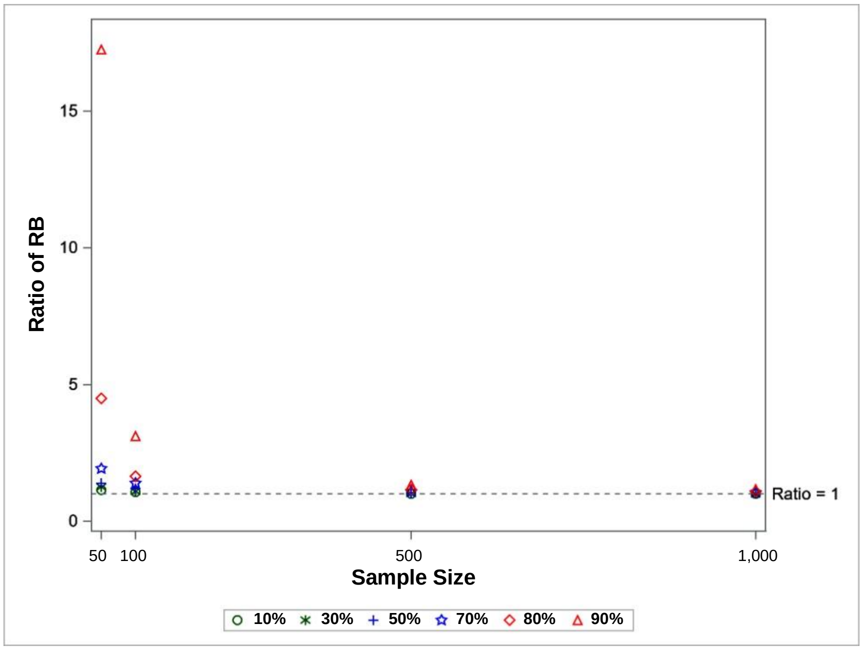


**Figure 3.5 Ratio of relative biases in standard errors for SRS samples for Z3.**

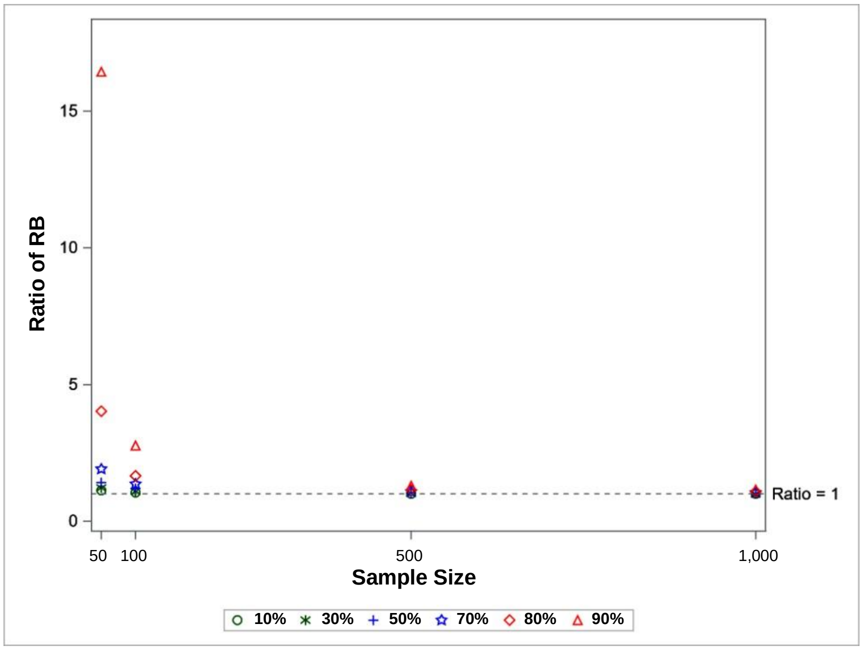


**Figure 3.6 Ratio of relative biases in standard errors for SRS samples for Z8.**

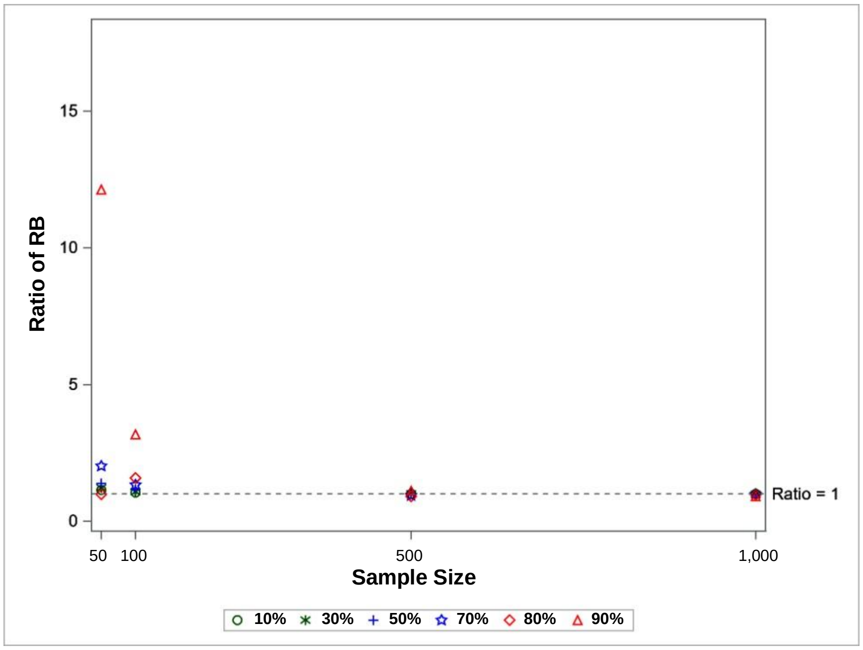


**Figure 3.7 Ratio of relative biases in standard errors for PPS samples for Z3.**

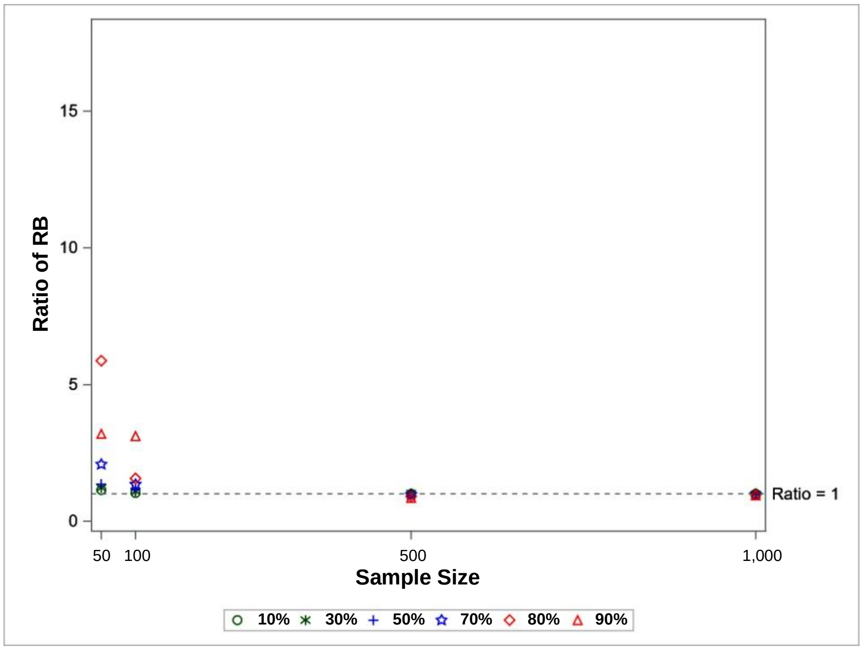


**Figure 3.8 Ratio of relative biases in standard errors for PPS samples for Z8.**

Table 3.1 displays the first quartile, median, and third quartile for ratio of RBs in point estimates and standard errors for sample size 50. Results for variable $Z_3$ for 10% and 90% censoring are reported in the table. We observed that for all variables, the first and third quartiles for ratio of RBs do not contain 1 when the sample size is small and the percentage of censoring is high. However, as expected, for large samples and a small number of censored observations, the difference in RBs between the scaled and unscaled weights is small.

**Table 3.1**
**Ratio of RBs in point estimates and standard errors for sample size 50 (variable $Z_3$)**

| Design | Ratio of RBs in Point Estimates | | | | | |
|---|---|---|---|---|---|---|
| | 90% Censored | | | 10% Censored | | |
| | First Quartile | Median | Third Quartile | First Quartile | Median | Third Quartile |
| SRS | 1.81 | 4.38 | 7.37 | 1.00 | 1.03 | 1.06 |
| PPS | 3.36 | 5.73 | 11.54 | 0.99 | 1.03 | 1.08 |
| | Ratio of RBs in Standard Errors | | | | | |
| Design | 90% Censored | | | 10% Censored | | |
| | First Quartile | Median | Third Quartile | First Quartile | Median | Third Quartile |
| SRS | 9.03 | 17.26 | 40.87 | 1.03 | 1.15 | 1.33 |
| PPS | 5.57 | 12.13 | 29.92 | 1.00 | 1.15 | 1.33 |

## 3.2 An application using NHEFS

We studied the time to get a heart attack and its relation to blood cholesterol and smoking using a data set from NHEFS.

The NHEFS is a national longitudinal survey in the United States that is used to determine the relationships between clinical, nutritional, and behavioral factors; to determine hospital utilizations; and to monitor changes in risk factors for an initial cohort that represents the NHANES I population. A cohort of size 14,407 was selected for the NHEFS. Vital and tracing status data, interview data, health care facility stay data, and mortality data from 1987 are available for public use. For more information about the survey and the data sets used in this section, see the Centers for Disease Control and Prevention’s website (https://www.cdc.gov/).

We used 4,673 observations from 1987 NHEFS public-use interview data to study the occurrence of first heart attack for the 1987 survey population and its relation to blood cholesterol and smoking. The following variables are used:

- Stratum, the stratum identification.
- ObservationWeight, the sampling weight associated with each observation unit.
- PSU, the primary sampling unit identification.
- Age, the event-time variable, defined as follows:
  - age of the subject when the first heart attack was reported for subjects who reported a heart attack,
  - age of the subject as reported in the interview for subjects who never reported a heart attack.
- HeartAttack, the heart attack indicator (1 = heart attack reported).

- Income, household income standardized to mean zero.
- HighBloodChol, the indicator that a subject has a high or low blood cholesterol level.
- Smoker, subject’s smoking habit (1 = current, 2 = former, –1 = non-smoker).
- Race, the race of the subject (1 = black, 2 = white, 3 = other).
- Gender, the gender of the subject.

The SURVEYPHREG procedure in SAS/STAT (Mukhopadhyay, 2010) is used to fit a proportional hazards regression model for age on income, blood cholesterol, smoking habit, race, gender, and race and gender interaction. Heart attack is used as the censored indicator. Observation weights range from 1,164 to 121,040 with mean 16,036.51, median 12,321, and coefficient of variation 74.35. The subjects are divided into 644 clusters and 35 strata.

PROC SURVEYPHREG is used in this section instead of PROC PHREG because the NHEFS uses complex survey design involving stratification, clustering, and unequal weights. PROC SURVEYPHREG supports STRATA, CLUSTER, and WEIGHT statements to account for stratification, clustering, and unequal weights, respectively. In addition, PROC SURVEYPHREG supports both Taylor series linearization and jackknife variance estimation methods for survey data (Mukhopadhyay, 2010). We used the jackknife variance estimation method for this study. SAS statements to fit this model are given in Appendix 2.

The 4,673 subjects in the sample represent almost 74.9 million individuals in the 1987 study population. Among all the subjects, 213 subjects reported at least one heart attack, and the other 4,460 subjects are considered to be censored. The 213 event observations in the sample represent an estimated 3.2 million population units, and the 4,460 censored observations in the sample represent an estimated 71.7 million population units. There are 95.44% observations in the sample that have not reported a heart attack which estimates 95.68% individuals in the population (Table 3.2) without a heart attack.

**Table 3.2**
**Number of censored and uncensored observations and their sum of weights**

| | Total | Event | Censored | Percent Censored |
|---|---|---|---|---|
| **Number of Observations** | 4,673 | 213 | 4,460 | 95.44 |
| **Sum of Weights** | 74,938,614 | 3,239,653 | 71,698,961 | 95.68 |

Without the Firth penalty, the Newton-Raphson optimization converges by satisfying the relative gradient convergence criterion (GCONV = 1E–8), but coefficients for the variables Smoker and Race do not converge. The coefficients for Smoker = 2 are 7.47, 10.87, and 11.83; and the coefficients for Race = 1 are 7.55, 10.95, and 11.17 in the last three iterations, respectively. This phenomenon is very common when you have a monotone likelihood (see Table 1.1). Among 644 replicate samples

(= 644 PSUs), monotone likelihood is observed in 542 replicates. Firth's penalized likelihood is a good alternative when you encounter monotone likelihoods.

We use the FIRTH option in PROC SURVEYPHREG (see "The SURVEYPHREG Procedure" in SAS Institute Inc. (2018)) to maximize Firth's penalized likelihood. The FIRTH option in PROC SURVEYPHREG uses the scaled weights. The penalized likelihood optimization converges with GCONV = 1E–8, along with reasonable convergence in all coefficients. Convergence is also achieved in all 644 replicate samples with the Firth penalty.

Table 3.3 displays the estimated hazards ratios along with their 95% Wald confidence intervals for blood cholesterol levels and smoking. In the 1987 study population, the estimated hazard of having a heart attack for a subject with low blood cholesterol is 0.6 times the estimated hazard of having a heart attack for a subject with high blood cholesterol. Because the 95% confidence interval does not contain 1, it is reasonable to conclude that the hazard of having a heart attack for a subject with low blood cholesterol is significantly lower than the hazard of having a heart attack for a subject with high blood cholesterol after adjusting for smoking, race, and other regressors in the 1987 study population.

The estimated hazard ratios for nonsmokers, current smokers, and former smokers are 0.59, 0.64, and 1.1, respectively. The estimated hazard for nonsmokers to have a heart attack is lower than the estimated hazard for current or former smokers. However, we do not have enough evidence to conclude that hazard ratios for smoking are significantly different at the 95% level after adjusting for blood cholesterol, race, and other regressors in the 1987 study population.

**Table 3.3**
**Hazard ratios for blood cholesterol and smoking, and their 95% Wald confidence intervals**

| | | Confidence Limit | |
|---|---|---|---|
| | **Point Estimate** | **Lower** | **Upper** |
| **HighBloodChol 0 vs 1** | 0.643 | 0.469 | 0.882 |
| **Smoker -1 vs 1** | 0.590 | 0.259 | 1.345 |
| **Smoker -1 vs 2** | 0.641 | 0.361 | 1.140 |
| **Smoker 1 vs 2** | 1.087 | 0.359 | 3.290 |

# 4 Summary

Firth's penalized likelihood is useful for obtaining maximum likelihood estimates from a monotone likelihood from proportional hazards regression models. We proposed a weight scaling method and demonstrated that Firth's penalized likelihood using the scaled weights have some desirable properties for complex surveys. A simulation study shows that estimated biases in point estimates and standard errors using the scaled weights are lower than estimated biases using the unscaled weights. Although Firth's penalized likelihood produces "good" estimates in most simulated data sets, there are a few data sets in which the Firth penalty failed to produce "good" convergences. The Firth penalized likelihood that uses scaled weights successfully corrected for a monotone likelihood when we estimated hazard rates for heart attacks using a data set from the NHEFS. Although the numeric results are quite encouraging, further research is needed to derive asymptotic distributions of the estimators obtained by using Firth's penalized likelihood.

We recommend the unpenalized likelihood when convergence is not an issue, but we recommend Firth's penalized likelihood using the scaled weights when a monotone likelihood is encountered in fitting proportional hazards regression models for complex surveys.

## Acknowledgements

I am grateful to Ying So, Randy Tobias, and Ed Huddleston at SAS Institute Inc. for their valuable assistance in the preparation of this article. I would also like to thank the two anonymous referees and the associate editor for their constructive suggestions.

## Appendix 1

### Consistency of the Firth penalized likelihood estimator

The estimators in Section 2 are defined as the solution to a system of equations that are constructed by using the score functions from proportional hazards regression models. In this appendix, we show that under certain regularity conditions these estimators are design consistent. Properties of estimators that are solutions to a set of estimating equations are well studied in the survey literature. For example, see Binder (1983), Godambe and Thompson (1986), and Fuller (2009, Section 1.3.4).

However, the estimating equations for proportional hazards regression models are more complex than the estimating equations for generalized linear models because the score functions involve weighted sums over the sampled units. Binder (1992) and Lin (2000) showed that the estimators obtained by solving the estimating equations for proportional hazards regression models are consistent. In this appendix, we follow arguments similar to those of Lin (2000) and Andersen and Gill (1982).

Several technical assumptions are necessary to show that the point estimates are consistent. We need assumptions about the estimating equations, the finite population, and the sample design – to whit:

- The functions defining the estimating equations should be smooth and convex.
- The finite population should be such that the moments for population quantities that are used in defining the estimating equations exists.
- The sample design should be such that the Narain-Horvitz-Thompson (NHT) estimators (Rao, 2005) for the population totals are well behaved.

All these assumptions are common in the sample survey literature; for example, see Fuller (2009). The score functions for proportional hazards regression models involve ratios of means of exponential functions that are infinitely differentiable.

Let $\mathcal{U}_N$ and $\mathcal{F}_N$ denote, respectively, the index set and values for the $N^{\text{th}}$ finite population in a sequence of populations indexed by $N,$ and let $A_N$ be a sample of size $n$ from $\mathcal{U}_N.$ To study large sample properties for sample-based estimators, we assume sequences of population and samples such that $N \rightarrow \infty$ and $(N - n) \rightarrow \infty,$ keeping the sampling fraction, $\frac{n}{N},$ fixed.

Assume $\mathcal{F}_N = \{(t_i, \Delta_i, \mathrm{Z}_i(\cdot))\}_{i=1}^{N}$ is an independent random sample of size $N$ from the joint distribution of $(T, \Delta, \mathrm{Z}(\cdot)),$ where $t$ is the failure time or the censoring time, whichever is less; $\Delta = 1$ if the failure time is less than the censoring time and 0 otherwise; and $Z(\cdot)$ is a vector of possibly time-varying explanatory variables.

Let $\boldsymbol{\beta}$ be a set of regression parameters for the superpopulation that is defined by the joint distribution of $(T, \Delta, Z(\cdot)).$ Let $\boldsymbol{\beta}_N$ be a set of finite population parameters obtained by solving the estimating equations when all $N$ units in the population are observed, and let $\hat{\boldsymbol{\beta}}_N$ be an estimator of $\boldsymbol{\beta}_N$ that is obtained by solving the weighted estimating equations by using only the sampled units. Our objective is to show that $\hat{\boldsymbol{\beta}}_N$ approaches $\boldsymbol{\beta}_N$ and that they both approach $\boldsymbol{\beta}$ as the sample size and population size increase.

Consider the estimating equations that correspond to Firth's penalized likelihood described in Section 2. For simplicity, we write these equations when there are no tied events. To further simplify notation, we write each component of the estimating equations separately. The finite population parameters, $\boldsymbol{\beta}_N,$ are a solution to the penalized partial likelihood score function, $U_N(\boldsymbol{\beta}) = (U_{N,1}(\boldsymbol{\beta}), \ldots, U_{N,P}(\boldsymbol{\beta}))',$ where

$$\begin{aligned} U_{N,p}(\boldsymbol{\beta}) = N^{-1} \sum_{i \in \mathcal{U}_N} \Delta_i \Biggl[ & \mathbf{Z}_i(t_i) - \frac{S_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \\ & + 0.5 \operatorname{tr} \Biggl( \Biggl[ N^{-1} \sum_{i \in \mathcal{U}_N} \Delta_i \Biggl\{ \frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \left( \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left( \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \Biggr\} \Biggr]^{-1} \\ & \qquad \Biggl\{ \left( \frac{Q_p^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \\ & \qquad - \left( \frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left( \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \\ & \qquad - \left( \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left( \frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \Biggr\} \Biggr) \Biggr] \end{aligned}$$

where

$$S^{(a)}(\boldsymbol{\beta}, t) = N^{-1}\sum_{i\in\mathcal{U}_N} I(t_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))[\mathbf{Z}_i(t)]^{\otimes a}$$

$$Q_p^{(a)}(\boldsymbol{\beta}, t) = N^{-1}\sum_{i\in\mathcal{U}_N} I(t_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))\, Z_{i,p}(t)[\mathbf{Z}_i(t)]^{\otimes a}$$

and where $a = 0, 1, 2$; $Z_i(t) = \left(Z_{i,1}(t), \ldots, Z_{i,p}(t)\right)'$; $\mathbf{S}^{(1)}(\boldsymbol{\beta}, t) = (S_1^{(1)}(\boldsymbol{\beta}, t), \ldots, S_P^{(1)}(\boldsymbol{\beta}, t))'$; $\mathrm{tr}(\cdot)$ denotes the trace of a matrix; $I(\cdot)$ denotes the indicator function; $p = 1, 2, \ldots, P$; and $P$ is the number of regression parameters. Note that $S^{(a)}(\boldsymbol{\beta}, t)$ and $Q_p^{(a)}(\boldsymbol{\beta}, t)$ depend on $N$, although the notation does not reflect this for reasons of simplicity.

In defining the score function for the penalized likelihood, we assume that the information matrix for the finite population, $I_N(\boldsymbol{\beta}, t)$, is always positive definite.

However, in any realistic situation, not all units in the finite population are available. Let a sample $A_N$ be selected by using a probability design that assigns a nonzero selection probability, $\pi_i$, to every unit in the population. Let $w_i = \pi_i^{-1}$ be the design weight. A sample-based estimator, $\hat{\boldsymbol{\beta}}_N$, is obtained by solving the estimated penalized partial likelihood score equations. Assuming that $N$ is known, a sample-based estimator for $U_{N,p}(\boldsymbol{\beta})$ is

$$\begin{aligned}
\hat{U}_{N,p}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in A_N} w_i\Delta_i \Bigg[ & \mathbf{Z}_i(t_i) - \frac{\hat{S}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \\
& + 0.5\mathrm{tr}\Bigg(\Bigg[N^{-1}\sum_{i\in A_N} w_i\Delta_i \Bigg\{\frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\}\Bigg]^{-1} \\
& \Bigg\{\left(\frac{\hat{Q}_p^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right) \\
& - \left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)' \\
& - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\}\Bigg)\Bigg]
\end{aligned}$$

where

$$\hat{S}^{(a)}(\boldsymbol{\beta}, t) = N^{-1}\sum_{i\in A_N} w_i I(t_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))[\mathbf{Z}_i(t)]^{\otimes a}$$

$$\hat{Q}_p^{(a)}(\boldsymbol{\beta}, t) = N^{-1}\sum_{i\in A_N} w_i I(t_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))Z_{i,p}(t)[\mathbf{Z}_i(t)]^{\otimes a}$$

are the NHT estimators for $S^{(a)}(\boldsymbol{\beta}, t)$ and $Q_p^{(a)}(\boldsymbol{\beta}, t)$, respectively.

Because $\hat{S}^{(a)}(\boldsymbol{\beta}, t)$ and $\hat{Q}_p^{(a)}(\boldsymbol{\beta}, t)$ use weighted sums over sampled units, we need techniques defined in Lin (2000) to study large sample properties of these estimators. Define $G_i(t) = \Delta_i I(t_i \leq t)$, $G(t) = N^{-1}\sum_{i\in\mathcal{U}_N} G_i(t)$, and $\hat{G}(t) = \sum_{i\in A_N} w_i G_i(t)$. Then the finite population score functions can be written using stochastic integration,

$$\begin{aligned}U_{N,p}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in\mathcal{U}_N}\int_0^\infty \Bigg[\mathbf{Z}_i(t_i) - \frac{S_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \\ + 0.5\mathrm{tr}\Bigg(\Bigg[N^{-1}\sum_{i\in A_N}\int_0^\infty\Bigg\{\frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\}dG(t_i)\Bigg]^{-1} \\ \Bigg\{\left(\frac{Q_p^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right) \\ - \left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)' \\ - \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\}\Bigg)\Bigg]dG_i(t)\end{aligned}$$

and the sample-based score functions are

$$\begin{aligned}\hat{U}_{N,p}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in\mathcal{U}_N}\int_0^\infty I(i\in A_n)w_i\Bigg[\mathbf{Z}_i(t_i) - \frac{\hat{S}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \\ + 0.5\mathrm{tr}\Bigg(\Bigg[N^{-1}\sum_{i\in\mathcal{U}_N}\int_0^\infty I(i\in A_N)w_i \\ \Bigg\{\frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\}dG_i(t)\Bigg]^{-1} \\ \Bigg\{\left(\frac{\hat{Q}_p^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right) \\ - \left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)' \\ - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\}\Bigg)\Bigg]dG_i(t).\end{aligned}$$

Note that the quantities $S^{(a)}$ and $Q_p^{(a)}$ are simply means over finite population quantities. Define the limits of these means as follows:

$$\begin{aligned}
\mathbf{s}^{(a)}(\boldsymbol{\beta}, t) &:= \lim_{N\to\infty} \mathbf{S}^{(a)}(\boldsymbol{\beta}, t) \\
&= \lim_{N\to\infty} N^{-1} \sum_{i\in\mathcal{U}_N} I(t_i \geq t) \exp(\boldsymbol{\beta}'\mathbf{Z}_i(t)) [\mathbf{Z}_i(t)]^{\otimes a} \\
q_p^{(a)}(\boldsymbol{\beta}, t) &:= Q_p^{(a)}(\boldsymbol{\beta}, t) \\
&= N^{-1} \sum_{i\in\mathcal{U}_N} I(t_i \geq t) \exp(\boldsymbol{\beta}'\mathbf{Z}_i(t)) Z_{i,p}(t) [\mathbf{Z}_i(t)]^{\otimes a} \\
g(t) &:= \lim_{N\to\infty} N^{-1} \sum_{i\in\mathcal{U}_N} G_i(t) \\
\alpha &:= \lim_{N\to\infty} N^{-1} \sum_{i\in\mathcal{U}_N} \int_0^\infty \mathbf{Z}_i(t)\, dG_i(t).
\end{aligned}$$

Thus the finite population score function, $U_{N,p}(\boldsymbol{\beta})$, converges to the superpopulation score function $u_{N,p}(\boldsymbol{\beta})$, where

$$\begin{aligned}
u_{N,p}(\boldsymbol{\beta}) = \alpha &- \int_0^\infty \frac{s_p^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}\, dg(t) \\
&+ 0.5\text{tr}\left( \int_0^\infty \left[ \int_0^\infty \left\{ \frac{s^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \left( \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right) \left( \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right)' \right\} dg(t) \right]^{-1} \right. \\
&\qquad \left\{ \left( \frac{q_p^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \frac{q_p^{(0)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \frac{s^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right) \right. \\
&\qquad - \left( \frac{\mathbf{q}_p^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \frac{q_p^{(0)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right) \left( \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right)' \\
&\qquad \left. \left. - \left( \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right) \left( \frac{\mathbf{q}_p^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \frac{q_p^{(0)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right)' \right\} dg(t) \right).
\end{aligned}$$

Now assume that the population quantities, $\mathbf{Z}_i$, that are used to define the score functions have finite moments and the sequence of sample designs is such that any smooth functions of NHT estimators are consistent. Because $U_N(\boldsymbol{\beta})$ is a smooth function of population totals, and each total is estimated by using a NHT estimator, $\hat{U}_N(\boldsymbol{\beta})$ is design-consistent for $U_N(\boldsymbol{\beta})$. That is, $\left(U_N(\boldsymbol{\beta}) - \hat{U}_N(\boldsymbol{\beta})\right) \big| \mathcal{F}_N = o(1)$. Therefore, by using arguments similar to Lin (2000) and Andersen and Gill (1982), it can be shown that $\boldsymbol{\beta}_N$ and $\hat{\boldsymbol{\beta}}_N$ converge to the same limit.

Because $n/N$ is fixed, $\sum_{A_N} w_i$ is the NHT estimator for $N$, and $\hat{U}(\boldsymbol{\beta})$ is a consistent estimator (not necessarily unbiased) of 0, both $\hat{U}(\boldsymbol{\beta})$ and $(n/N)\hat{U}(\boldsymbol{\beta})$ converge to the same limit with the same order of convergence. It is straightforward to show that $(n/N)\hat{U}(\boldsymbol{\beta})$ and $\left(n/\sum_{A_N} w_i\right)\hat{U}(\boldsymbol{\beta})$, the estimating equations that use the scaled weights, have the same expectation.

# Appendix 2

## SAS program to obtain the Firth penalized likelihood estimates

The SAS statements at the end of this section fit a proportional hazards regression model using the scaled weights in Firth’s penalized likelihood. The PROC statement invokes the procedure, and the VARMETHOD = JK option requests the jackknife variance estimation method. You can also specify VARMETHOD = TAYLOR, VARMETHOD = BRR, or VARMETHOD = BOOT to request the Taylor series linearized, balanced repeated replication, or bootstrap replication variance estimation method, respectively. The DETAILS sub-option of the VARMETHOD = JK option prints estimates from each replicate sample along with the convergence status. The WEIGHT statement specifies the sampling weights, the STRATA statement specifies the strata, and the CLUSTER statement specifies the PSUs. The MODEL statement specifies the analysis model. The FIRTH option in the MODEL statement requests Firth’s penalized likelihood. The two HAZARDRATIO statements requests hazard ratios for blood cholesterol and smoking, respectively. The ODS OUTPUT statement stores replicate estimates and convergence status from each replicate in the SAS data set RepEstimatesFirth. This data set is useful for checking the convergence status of every replicate sample.

```
proc surveyphreg data = NHEFS varmethod=jk (details);
      class        Gender HighBloodChol Race Smoker;
      weight       ObservationWeight;
      strata       Stratum;
      cluster      PSU;
      model        EventTime*HeartAttack(2)  =  Income  HighBloodChol
                   Smoker Race Gender Race*Gender / firth;
      hazardratio HighBloodChol;
      hazardratio Smoker;
      ods output  repestimates=RepEtimatesFirth;
run;
```

# References

Andersen, P.K., and Gill, R.D. (1982). Cox’s regression model counting process: A large sample study. *Annals of Statistics*, 10, 1100-1120.

Bender, R., Augustin, T. and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24, 1713-1723.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D.A. (1990). Fitting Cox's proportional hazards models from survey data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 342-347.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika*, 79, 139-147.

Binder, D.A., and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89(427),1035-1043.

Boudreau, C., and Lawless, J.F. (2006). Survival analysis based on the proportional hazards model and survey data. *Canadian Journal of Statistics*, 34, 203-216.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89-99.

Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*. Chichester, UK: John Wiley & Sons, Inc.

Chambless, L.E., and Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics - Theory and Methods*, 14, 1377-1392.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, (with discussion), 34, 187-220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27-38.

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, *Series C*, 37, 117-132.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54, 127-138.

Heinze, G. (1999). *The Application of Firth's Procedure to Cox and Logistic Regression*. Technical Report 10, updated January 2001, Department of Medical Computer Sciences, Section of Clinical Biometrics, University of Vienna.

Heinze, G., and Schemper, M. (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, 51, 114-119.

Heinzel, G., Rüdiger, A. and Schilling, R. (2002). *Spectrum and Spectral Density Estimation by the Discrete Fourier Transform (DFT), Including a Comprehensive List of Window Functions and Some New Flat-Top Windows*. Tech. rep., Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Teilinstitut Hannover.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art*, (Eds., J.P. Klein and P.K. Goel), Dordrecht, Netherlands: Kluwer Academic, 237-247.

Lin, D.Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika*, 87, 37-47.

Lin, D.Y., and Wei, L.J. (1989). The robust inference for the proportional hazards model. *Journal of the American Statistical Association*, 84, 1074-1078.

Mukhopadhyay, P.K. (2010). Not hazardous to your health: Proportional hazards modeling for survey data with the SURVEYPHREG procedure. *Proceedings of the SAS Global Forum 2010 Conference*. Cary, NC: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings10/254-2010.pdf.

Mukhopadhyay, P.K., An, A.B., Tobias, R.D. and Watts, D.L. (2008). Try, try again: Replication-based variance estimation methods for survey data analysis in SAS 9.2. *Proceedings of the SAS Global Forum 2008 Conference*. Cary, NC: SAS Institute Inc. http://www2.sas.com/proceedings/forum2008/367-2008.pdf.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 2, 117-138. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9040-eng.pdf.

SAS Institute Inc. (2018). *SAS/STAT 15.1 User's Guide*. Cary, NC: SAS Institute Inc. http://go.documentation.sas.com/?docsetId=statug&docsetTarget=titlepage.htm&docsetVersion=15.1&locale=en.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2nd Ed. New York: Springer.

Vol. 46, No. 2, pp. 243-264
Statistics Canada, Catalogue No. 12-001-X

# Probability-proportional-to-size ranked-set sampling from stratified populations

**Omer Ozturk**[1]

## Abstract

This paper constructs a probability-proportional-to-size (PPS) ranked-set sample from a stratified population. A PPS-ranked-set sample partitions the units in a PPS sample into groups of similar observations. The construction of similar groups relies on relative positions (ranks) of units in small comparison sets. Hence, the ranks induce more structure (stratification) in the sample in addition to the data structure created by unequal selection probabilities in a PPS sample. This added data structure makes the PPS-ranked-set sample more informative then a PPS-sample. The stratified PPS-ranked-set sample is constructed by selecting a PPS-ranked-set sample from each stratum population. The paper constructs unbiased estimators for the population mean, total and their variances. The new sampling design is applied to apple production data to estimate the total apple production in Turkey.

**Key Words:** PPS sampling; Stratified PPS; Ranked-set sample; Sample allocation.

# 1 Introduction

In survey sampling studies, selection of a sampling design depends on the structure of the population. In this paper, we consider a population structure having two main features. It must contain a size variable $X$ and a variable of interest $Y$. The values of the size variable should be approximately proportional to the values of the $Y$-variable, and the values of the $X$-variable should be available for all population units prior to sampling. The second feature of the population structure is that a small percentage of population units should produce extreme values in both the $Y$- and $X$-variables with different proportionality constants. These units usually produce larger means and variances than the rest of the units in the population in both the $Y$- and $X$-variables. This population structure is very common in practice. In agricultural sampling, a farm population in a state or country may contain two variables, the crop production $Y$ and the farm size $X$ in acres. Farms can be divided into two groups, the farms that have small/normal sizes and the mega-farms that have extremely large values in the $X$- and $Y$-variables. The percentage of the mega-farms would be small, but they may have larger means and variances in the $Y$- and $X$-variables and the proportionality constant between the $Y$- and $X$-values may be larger.

The population structure in the Monthly Retail Trade Survey performed by the United States Census Bureau would be another example. In this case, the population is defined by the business establishments that have Employer Identification Numbers (EINs). The Census Bureau uses a very complex design in which the previous years' annual revenues are used to construct a size variable. The structure of the population fits the setting we consider in this paper. Revenues from the previous years would be approximately proportional to the current revenues. The revenues for most of the businesses would take typical values, while revenues of a certain percentage of businesses would be extremely large, producing a

[1] Omer Ozturk, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH, 43210, U.S.A. E-mail: omer@stat.osu.edu.

larger mean and variance and a different proportionality constant. Different proportionality constant may happen because large businesses may be more/less productive than the rest of the businesses.

We provide a third example using apple production data in Turkey in 2002. The data set was collected by the Turkish Statistical Institute and reported in Kadilar and Cingi (2003) and Ozturk and Bayramoglu-Kavlak (2018). It contains two variables, apple production $(Y)$ (in 1,000kg) and the numbers of apple trees $(X)$ in townships (or localities). The sampling units are the 851 localities in the data set. The $X$-values in all townships are available in the sampling frame prior to sampling. Figure 1.1 provides the scatter plot of the $Y$- and $X$-values, where we see that the value of the $Y$-variable is an increasing function of the $X$-value. We also observe that red-colored points marked with "$X$" in the plot have large values for both $X$- and $Y$-variables and their proportionality constant is different from the other points. Hence, this population fits into our population structure.

The apple production data have additional structures. The entire population is stratified into seven different geographical regions: Marmara, Aegean, Mediterranean, Central Anatolia, Black Sea, Eastern Anatolia and Southeastern Anatolia. These regions have different climate patterns and apple production changes significantly from one region to another. The extreme observations marked with "$X$" in Figure 1.1 come from the Marmara, Aegean, Mediterranean and Central Anatolia regions. This is a natural setting to construct a PPS-ranked-set sample from each sub-population.

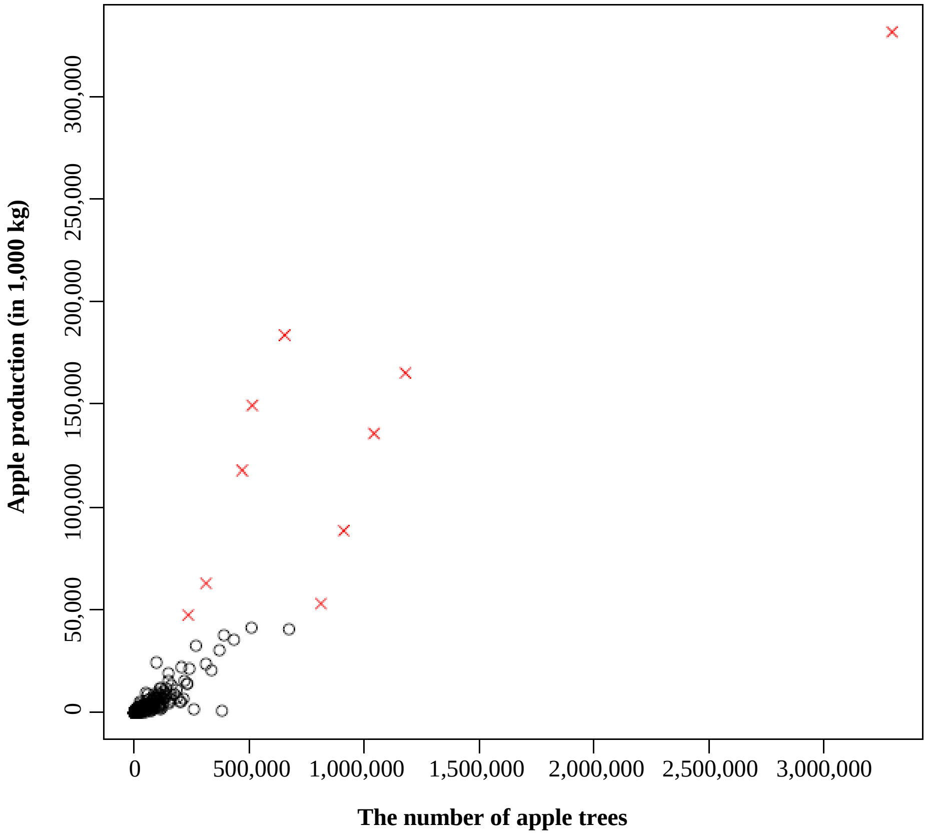


**Figure 1.1 Scatter plot of apple production.**

The variable $X$ in our population setting provides information about the relative size of the units in the entire population. Since the variable $Y$ is approximately proportional to the variable $X$, the size of the unit indicates the importance of its contribution to the variable of interest $Y$. Hence, important units should have higher probability of being included in the sample. Probability-proportional-to-size (PPS) sampling deliberately imposes higher selection probabilities for important units to produce unbiased and highly efficient estimators for the population mean and/or total. The main contribution of this paper is to introduce a new sampling design which combines the ranking information in a ranked-set sample (RSS) with the advantage of unequal selection probabilities as used in a PPS sample.

A typical PPS sample contains the triplets $(Y_i, \pi_i, u_i)$, $i = 1, \ldots, n$, where $Y_i$ and $\pi_i$ are the value of $Y$ and the selection probability of the unit $u_i$ for each draw under sampling with replacement selection. Though PPS sampling can be done without replacement, here all references to PPS sampling refers to sampling with replacement. Readers may refer to Thompson (2002, page 53) for the details of PPS sampling. In a PPS sample, the size variable is not necessarily used directly in the construction of the estimators. On the other hand, the values of the $X$-variable are available for all population units even for the units not included in the PPS sample. Hence the $X$-variable could help us to borrow additional information from a comparison set of $H$ unmeasured units.

For the construction of a typical data point $(Y_i, \pi_i, u_i)$ in a PPS-ranked-set sample, we select $H$ units from the population using PPS sampling with replacement to form a comparison set $\{u_1, \ldots, u_H\}$. We rank these units without measurement based on the $X$-variable with no additional cost. We then measure the value of the $Y$-variable for only one unit, the unit having rank $R_i$, obtaining $(Y_{[R_i]}, \pi_{[R_i]}, u_{[R_i]})$, where $Y_{[R_i]}$ and $\pi_{[R_i]}$ are the value of the $Y$-variable and the selection probability of the unit $u_{[R_i]}$ at each draw. A data point in the comparison set, $(Y_{[R_i]}, \pi_{[R_i]}, u_{[i]})$, provides more information than a data point in a PPS sample, $(Y_i, \pi_i, u_i)$, since the rank $R_i$ borrows additional information from the other $H-1$ unmeasured units in the comparison set. In this paper, we use this idea to construct a PPS-ranked-set sample that is more informative than a PPS sample. The details of this sampling procedure will be provided in Section 2.

The position information is used in a slightly different context in ranked-set sample (RSS) and judgment-post-stratified (JPS) sampling designs to borrow information from the unmeasured population units. Construction of a ranked-set sample of size $n$ requires one to determine two integers $d$ and $H$, $n = dH$, where $H$ and $d$ are the set and cycle sizes, respectively. The set size $H$ controls the amount of information that can be borrowed from the units in comparison sets. The cycle size $d$ is used to increase the total sample size in a RSS. Once $H$ and $d$ are chosen, one then selects $nH$ units from the population and partitions them into $n$ disjoint comparison sets, each having $H$ units. Units in each comparison set are ranked without measurement using the $X$-variable and the value of the $Y$-variable $(Y_{[h]j};\ j = 1, \ldots, d)$ associated with the $h^{\text{th}}$ ranked $X$ is measured in $d$ different comparison sets, $h = 1, \ldots, H$. The measured values $Y_{[h]j}$, $h = 1, \ldots, H$; $j = 1, \ldots, d$, are called a ranked-set sample.

The construction of a JPS sample of size $n$ starts with a simple random sample of size $n$ and measures all of them, $Y_i$, $i = 1, \ldots, n$. For each measured unit in this sample, one then selects an additional $H-1$

units from the population to form a comparison set of size $H$. The rank $R_i$ of the measured unit $Y_i$ in each of these comparison sets is determined. The pairs of $(Y_i, R_i)$, $i = 1, \ldots, n$, constitute a JPS sample.

The RSS and JPS samples create induced order statistics for the $Y$-variable through the ranks of the $X$-variable in comparison sets. Hence, the random variable $Y_{[h]i}$ ($Y_i$ given that $R_i = h$ for the JPS sample) is stochastically smaller than the random variable $Y_{[h']j}$ ($Y_i$ given that $R_j = h'$) for $h \leq h'$. This stochastic ordering property induces an implicit stratification among the measured sample units. Efficiency improvements of RSS and JPS samples over a simple random sample can be anticipated from the partition of the total variation into between- and within-strata variation. For further details on these sampling designs, readers may refer to the review paper in Wolfe (2012) and references therein. Both RSS and JPS samples use the position information of the units in the comparison sets, but they do not completely use the information provided by the selection probabilities in a PPS sample. All units in comparison sets for RSS and JPS are selected with equal probabilities. Hence, they may not be appropriate for the population structure that we consider in this paper.

MacEachern, Stasny and Wolfe (2004) introduced the JPS design in an infinite population setting. In a finite population setting, constructions of the JPS and RSS samples depend on whether the comparison sets are selected with or without replacement. Patil, Sinha and Taillie (1995) considered an RSS design in a finite population, where none of the units in a comparison set is returned to the population prior to selection of the next comparison set. Deshpande, Frey and Ozturk (2006) expanded the RSS sampling design with three different without replacement selection policies and constructed nonparametric confidence intervals for population quantiles.

Probability sampling has also generated extensive research interest in RSS and JPS sampling. Al-Saleh and Samawi (2007), Ozdemir and Gokpinar (2007 and 2008), Gokpinar and Ozdemir (2010), Ozturk and Jafari Jozani (2013), Frey (2011) and Ozturk (2014) computed inclusion probabilities and constructed Horwitz-Thompson type estimators for the population mean and total based on a ranked-set sample. These research papers show that an RSS design yields a substantial amount of improvement in efficiency over the usual simple random sampling design. Ozturk (2016) developed estimators for the population mean based on a JPS sample, where he showed that the estimator needs a finite population correction factor similar to the one used in a simple random sample.

A few researchers have applied the RSS methodology to existing survey sampling designs. Muttlak and McDonald (1992) incorporated the RSS sampling design with a line intersect method. Sroka (2008) used it in stratified sampling by constructing an RSS sample from each stratum. Wang, Lim and Stokes (2016) considered the RSS design in a cluster randomized design with a mixed effect model, where the cluster effect is treated as random. They showed that use of RSS at the cluster level has much bigger impact on efficiency than using the RSS at the within-cluster level. Nematollahi, Salehi and Aliakbari Saba (2008) used the RSS design in a finite population setting only in the second stage of a two-stage sampling with replacement selection scheme. Since they use the RSS design only in the second stage with replacement, the efficiency improvement of their estimator with respect to a two-stage SRS sample estimator was minimal. Sud and Mishra (2006) also used a two-stage cluster sample with ranked set sampling design in a finite population setting under the assumption that the cluster population sizes are all equal. Ozturk (2019a) developed design based statistical inference for a two-stage clustered ranked-set sample in a finite population setting.

In this paper, we develop statistical inference for the PPS-ranked-set sampling design in a population setting where the values of the size variable are roughly proportional to the values of the variable of interest and a small percentage of population units produces large $X$- and $Y$-values with a different proportionality constant. We motivate the new sampling design using apple production data. Section 2 introduces the PPS-ranked-set sample in a finite population setting. It constructs unbiased estimators for the population mean, total and their variances. We show that the PPS-ranked-set sample estimator has smaller variance than a PPS sample estimator. Section 3 extends the PPS-ranked-set sample to a stratified population and constructs unbiased estimators for the population mean, total and their variances. Section 4 considers four different sample size allocation procedures to minimize the variance of the estimator under a cost model and different stratum population structures. Section 5 provides an efficiency comparison for the PPS-ranked-set sample estimator of the population mean with respect to other competing estimators. Section 6 illustrates the use of PPS-ranked-set sample data to estimate the apple production in Turkey. Section 7 provides some concluding remarks. All proofs are given in the Appendix.

# 2 Probability-proportional-to-size ranked-set sample

We consider a finite population of size $N$, $\mathcal{P}^N = \{u_1, \ldots, u_N\}$. Assume that each unit in the population possesses two characteristics $Y$ and $X$. The values of $X$ are available for all units and the values of $Y$ are roughly proportional to the values of $X$. A small percentage of the population units may produce extreme values for both $Y$- and $X$-variables with a different proportionality constant. Let $\pi_i$ be the probability that the unit $u_i$ is selected from $\mathcal{P}^N$, for each draw.

$\pi_i = P$ (unit $u_i$ is selected from $\mathcal{P}^N$ under sampling with replacement selection for each draw),

where $\pi_i$ is proportional to the size of $X$ for unit $u_i$. The values of the variable $Y$ in the population $\mathcal{P}^N$ are denoted by $y_1, \ldots, y_N$. The mean and variance of this population are defined as

$$\mu_N = \frac{1}{N} \sum_{k=1}^{N} y_k \text{ and } \sigma_N^2 = \frac{1}{N} \sum_{k=1}^{N} (y_k - \mu_N)^2 .$$

We first briefly introduce the notation for a PPS sample. Let $n$ be the sample size. We consider a PPS sample, $(Y_i, \pi_i, u_i)$, $i = 1, \ldots, n,$ constructed from population $\mathcal{P}^N$ under a sampling with replacement selection scheme with selection probability $\pi_i$. The probability mass function (PMF) and the cumulative distribution function (CDF) of the random variable $Y_i$ are given by

$$P(Y_i = y) = f(y) = \sum_{j=1}^{N} \pi_j I(y_j = y), \quad F(y) = \sum_{z=y_1}^{y} \sum_{j=1}^{N} \pi_j I(y_j = z).$$

We note that, since the sample units $u_i,\ i = 1, \ldots, n,$ are selected with replacement, the $Y_i,\ i = 1, \ldots, n,$ are independent and identically distributed. We now consider the order statistics $Y_{(h:\,n)}$ in a sample of size $n$. The CDF of the $h^{\text{th}}$ order statistics in a sample of size $n$ is given by

$$F_{(h:\,n)}(y) = \sum_{r=h}^{n} \binom{n}{r} F^{r}(y)\,\{1 - F(y)\}^{n-r}$$

$$f_{(h:\,n)}(y) = F_{(h:\,n)}(y) - F_{(h:\,n)}(y^{-}), \tag{2.1}$$

where $F_{(h:n)}(y^{-})$ is the left-limit at $y$.

We now construct a PPS-ranked-set sample that combines the ranking information in comparison sets with the information provided by the selection probabilities in a PPS sample. Let $H$ be the set size. Using a PPS sampling design, we select $H$ units from the population with replacement to form a comparison set $S = \{u_1, \ldots, u_H\}$ with selection probabilities, $\{\pi_{u_1}, \ldots, \pi_{u_H}\}$. Units in this set are ranked from the smallest to the largest based on the values of the size variable $X$, yielding $S^{*} = \{(Y_{[1]}, \pi_{[1]}, u_{[1]}, X_{(1)}), \ldots, (Y_{[H]}, \pi_{[H]}, u_{[H]}, X_{(H)})\}$, where $Y_{[h]}$ is the value of the $Y$-variable and $\pi_{[h]}$ is the selection probability of the unit $u_{[h]}$ that correspond to the $h^{\text{th}}$ smallest $X$ $(X_{(h)})$ in the set. The smallest ranked unit in set $S^{*}$ is selected and measured for the variable $Y, Y_{[1]},$ and its selection probability, $\pi_{[1]},$ is recorded. The remaining $H - 1$ units are not measured. They are used only to obtain the rank of $u_1$ based on ranking of the $X$ - measurements. We construct another comparison set using a PPS sample and rank the units based on the $X$-variable. This time, we measure the $Y$-variable on the unit that corresponds to the second smallest $X$-value and record its selection probability, $(Y_{[2]}, \pi_{[2]})$. We continue constructing comparison sets and measuring the $Y$-variables until we have the measurement from the unit that corresponds to the largest $X$-value, $(Y_{[H]}, \pi_{[H]})$. The measured values $(Y_{[h]}, \pi_{[h]}),\ h = 1, \ldots, H,$ are called a cycle. To increase the sample size to $n = Hd,$ the entire process is repeated for $d$ cycles. The measured values $Y_{[h]j}, \pi_{[h]j},\ h = 1, \ldots, H,\ j = 1, \ldots, d$ are called a PPS-ranked-set sample, where $Y_{[h]j}, \pi_{[h]j}$ are the $Y$-measurement and selection probability of the unit $u_{[h]}$ that corresponds to the $h^{\text{th}}$ smallest value of the $X$-variable in the set $h$ and cycle $j$. We refer to $Y_{[h]i}$ as the induced $h^{\text{th}}$ order statistic since its position is induced by the $X$-values in the set. We note that the induced order statistic $Y_{[h]i}$ and induced ordered unit $u_{[h]i}$ are defined in comparison sets with set size $H$. To simplify the notation we omit the set size $H$ and write $Y_{[h]i} = Y_{[h;H]i}, u_{[h]i} = u_{[h:H]i}$.

The PPS-ranked-set sample is illustrated in Table 2.1 for the set size $H = 3$ and the cycle size $d = 2$. For each cycle, the table contains three comparison sets (rows). Each set has three units. The units in each set are ranked based on the $X$-variable. The units on the diagonal (bold faced) are measured for the values of the $Y$-variables, and their selection probabilities are recorded. The last column contains the measured values of the units in the PPS-ranked-set sample. Each measured data point in Table 2.1 provides three pieces of information: (1) the value of $Y$, (2) the selection probability of the unit under sampling with replacement, and (3) the relative position of the unit in its comparison set. One can make an intuitive comparison between the PPS-ranked-set sample and other sampling designs in the literature. For example, a simple random sample provides the information in item (1), a ranked-set sample provides the information in items (1) and (3), and a PPS sample provides the information in items (1) and (2). We anticipate (and show in Theorem 1) that the PPS-ranked-set sample is more informative than all three of these sampling designs, and therefore has smaller variance, since it provides the information in items (1), (2) and (3).

**Table 2.1**
**Illustration of the PPS-ranked-set sample for the set size $H = 3$ and cycle size $d = 2$**

| Cycle | Set | Ranked units in comparison sets | Measurements |
|---|---|---|---|
| 1 | 1 | $\{\mathbf{Y}_{[1]1}, \boldsymbol{\pi}_{[1]1}, \mathbf{X}_{(1)1}\}, \{Y_{[2]1}, \pi_{[2]1}, X_{(2)1}\}, \{Y_{[3]1}, \pi_{[3]1}, X_{(3)1}\}$ | $(Y_{[1]1}, \pi_{[1]1})$ |
|  | 2 | $\{Y_{[1]1}, \pi_{[1]1}, X_{(1)1}\}, \{\mathbf{Y}_{[2]1}, \boldsymbol{\pi}_{[2]1}, \mathbf{X}_{(2)1}\}, \{Y_{[3]1}, \pi_{[3]1}, X_{(3)1}\}$ | $(Y_{[2]1}, \pi_{[2]1})$ |
|  | 3 | $\{Y_{[1]1}, \pi_{[1]1}, X_{(1)1}\}, \{Y_{[2]1}, \pi_{[2]1}, X_{(2)1}\}, \{\mathbf{Y}_{[3]1}, \boldsymbol{\pi}_{[3]1}, \mathbf{X}_{(3)1}\}$ | $(Y_{[3]1}, \pi_{[3]1})$ |
| 2 | 1 | $\{\mathbf{Y}_{[1]2}, \boldsymbol{\pi}_{[1]2}, \mathbf{X}_{(1)2}\}, \{Y_{[2]2}, \pi_{[2]2}, X_{(2)2}\}, \{Y_{[3]2}, \pi_{[3]2}, X_{(3)2}\}$ | $(Y_{[1]2}, \pi_{[1]2})$ |
|  | 2 | $\{Y_{[1]2}, \pi_{[1]2}, X_{(1)2}\}, \{\mathbf{Y}_{[2]2}, \boldsymbol{\pi}_{[2]2}, \mathbf{X}_{(2)2}\}, \{Y_{[3]2}, \pi_{[3]2}, X_{(3)2}\}$ | $(Y_{[2]2}, \pi_{[2]2})$ |
|  | 3 | $\{Y_{[1]2}, \pi_{[1]2}, X_{(1)2}\}, \{Y_{[2]2}, \pi_{[2]2}, X_{(2)2}\}, \{\mathbf{Y}_{[3]2}, \boldsymbol{\pi}_{[3]2}, \mathbf{X}_{(3)2}\}$ | $(Y_{[3]2}, \pi_{[3]2})$ |

We note that $Y_{[h]1}$ is not necessarily the same as the $Y$-value of the unit having the $h^{\text{th}}$ smallest $Y$-value $(Y_{(h)1})$ since its rank is induced based on the size variable $X$. The square brackets are used to denote the possibility of within-set ranking error. If there is no ranking error, the square brackets are replaced with round parentheses. In this case $Y_{(h)1}$ becomes the $h^{\text{th}}$ order statistic in a set of size $H$.

In a recent study, Ozturk (2019b) used the induced ranks post-experimentally in a PPS-judgment-post-stratified sample. The key difference between a PPS-ranked-set sample and a PPS-judgment-post-stratified sample is in the implementation of the ranking process. The ranks in a PPS-ranked-set sample are obtained prior to measurement of the $Y$-variable, but the ranks in a PPS-judgment-post-stratified sample are obtained post-experimentally after measuring the $Y$-variables in a PPS sample.

Throughout the paper, our ranking procedure satisfies the consistency requirement

$$\mathrm{HF}(y) = \sum_{h=1}^{H} F_{[h:H]}(y), \tag{2.2}$$

where $F_{[h:H]}(y)$ is the CDF of $Y_{[h]i}$. The proof of equation (2.1) is provided in Presnell and Bohn (1999). The consistency in the ranking process indicates that the same ranking procedure, however imperfect it might be, is applied in all comparison sets. Hence, the equality in equation (2.1) holds for ranking methods that use the size variable $X$.

We construct an estimator for the population mean from the PPS-ranked-set sample:

$$\bar{Y}_{\mathrm{PR},\,N} = \frac{1}{HdN} \sum_{h=1}^{H} \sum_{i=1}^{d} \frac{Y_{[h]i}}{\pi_{[h]i}}.$$

An estimator for the population total is given by $T_{\mathrm{PR},N} = N\bar{Y}_{\mathrm{PR},N}$. The standard PPS estimator, often referred to as the Hansen-Hurwitz estimator, with sample size $n = dH$ has the same form as the estimator $\bar{Y}_{\mathrm{PR}}$, but it does not use the ranking information:

$$\bar{Y}_{P,N} = \frac{1}{nN}\sum_{i=1}^{n}\frac{Y_i}{\pi_i}.$$

The variance of $\bar{Y}_{P,N}$ is given in standard text books to be (see, for example, Thompson, 2002, page 52)

$$\sigma^2_{\bar{Y}_{P,N}} = \mathrm{Var}\left(\bar{Y}_{P,N}\right) = \frac{1}{nN^2}\,\mathrm{Var}\left(\frac{Y_1}{\pi_1}\right) = \frac{1}{nN^2}\sum_{k=1}^{N}\pi_k\left(\frac{y_k}{\pi_k} - N\mu_N\right)^2.$$

**Theorem 1.** *Let $\left(Y_{[h]i}, \pi_{[h]i}\right)$, $h = 1, \ldots, H$, $i = 1, \ldots, d$ be a PPS-ranked-set sample from population $\mathcal{P}^N$. Under any consistent ranking scheme satisfying equation (2.1), the estimator $\bar{Y}_{\mathrm{PR},N}\left(T_{\mathrm{PR},N}\right)$ is unbiased for the population mean (total). Their variances are equal to $\sigma^2_{\bar{Y}_{\mathrm{PR},N}}$ and $\sigma^2_{T_{\mathrm{PR},N}}$*

$$\begin{aligned}
\sigma^2_{\bar{Y}_{\mathrm{PR},N}} &= \frac{1}{H^2dN^2}\sum_{h=1}^{H}\left\{\sum_{k=1}^{N}\frac{y_k^2}{\pi_k^2}f_{[h:H]}(y_k) - \left[\sum_{k=1}^{N}\frac{y_k}{\pi_k}f_{[h:H]}(y_k)\right]^2\right\}\\
&= \frac{1}{H^2dN^2}\sum_{h=1}^{H}\sigma^2_{[h:H]} = \sigma^2_{\bar{Y}_{p,N}} - \frac{1}{nHN^2}\sum_{h=1}^{H}\left(\mu_{[h:H]} - N\mu_N\right)^2 \le \sigma^2_{\bar{Y}_{p,N}},\\
\sigma^2_{T_{\mathrm{PR},N}} &= N^2\sigma^2_{\bar{Y}_{\mathrm{PR},N}},
\end{aligned}$$

*where $\mu_{[h:H]} = E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)$ and $\sigma^2_{[h:H]} = E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \mu^2_{[h:H]}$ are the mean and variance of $Y_{[h]i}/\pi_{[h]i}$.*

We note that the last two expected values in Theorem 1 are computed using the randomization distribution. Theorem 1 shows that the estimator $\bar{Y}_{\mathrm{PR},N}$ has always smaller variance than the variance of the mean of a PPS estimator as long as there is meaningful information to rank the sample units in a comparison set. For settings where the PPS sample is appropriate, ranking information would be available since the size variable $X$ is approximately proportional to the variable $Y$. Hence, it provides reasonably accurate ranking for the units in the comparison sets.

The probability mass function in Theorem 1, $f_{[h:H]}(y_k)$, is given for perfect ranking as in equation (2.1). Under imperfect ranking, $f_{[h:H]}(y_k)$ is the PMF of the induced order statistic $Y_{[h]1}$, and its form is not known. In the next theorem, we provide an unbiased estimator for $\sigma^2_{\bar{Y}_{\mathrm{PR},N}}$ $\left(\sigma^2_{T_{\mathrm{PR},N}}\right)$ regardless of the quality of the ranking information.

**Theorem 2.** *Let $\left(Y_{[h]i}, \pi_{[h]i}\right)$, $h = 1, \ldots, H$; $i = 1, \ldots, d$, be a PPS-ranked-set sample from population $\mathcal{P}^N$. Under any consistent ranking scheme satisfying equation (2.1), unbiased estimators of $\sigma^2_{\bar{Y}_{\mathrm{PR},N}}$ and $\sigma^2_{T_{\mathrm{PR},N}}$ are given by*

$$\begin{aligned}
\hat{\sigma}^2_{\bar{Y}_{\mathrm{PR},N}} &= \frac{1}{2d^2(d-1)H^2N^2}\sum_{h=1}^{H}\sum_{i=1}^{d}\sum_{j\ne i}^{d}\left\{\frac{Y_{[h]i}}{\pi_{[h]i}} - \frac{Y_{[h]j}}{\pi_{[h]j}}\right\}^2, \quad d > 1\\
\hat{\sigma}^2_{T_{\mathrm{PR},N}} &= N^2\hat{\sigma}^2_{\bar{Y}_{\mathrm{PR},N}}.
\end{aligned}$$

For moderately large sample sizes, we can use the normal approximation to provide approximate $(1-\alpha)\,100\%$ confidence intervals for the population mean and total, namely,

$$\bar{Y}_{\text{PR},\,N} \pm t_{n-H,\,\alpha/2}\hat{\sigma}^2_{\bar{Y}_{\text{PR},\,N}},\quad \bar{T}_{\text{PR},\,N} \pm t_{n-H,\,\alpha/2}\hat{\sigma}^2_{\bar{T}_{\text{PR},\,N}},$$

where $t_{n-H,\,a}$ is the $a^{\text{th}}$ upper quantile of a t-distribution having degrees of freedom $df = n - H.$ The $df = n - H$ is proposed to take into account the heterogeneity between judgment ranking classes. For smaller sample sizes, one can approximate the degrees of freedom using the Satterthwaite approximation.

We now investigate the efficiency of the PPS-ranked-set sample estimator using several populations that fit the structure presented in Section 1. The finite populations are generated using the model below.

I. For a fixed population size $N,$ generate the size variable $X$ from an exponential distribution with mean 100 and order these $N$ random numbers from the smallest to the largest, $x_{(1)} < \ldots < x_{(N)},$ where $x_{(i)}$ is the $i^{\text{th}}$ smallest value of the $X$-values.

II. Let $N^*$ be he largest integer such that $N^* \le N(1-\omega).$ Generate the $Y$-values from either

$$y_{[i]} = \begin{cases} x_{(i)} + \tau\varepsilon_i & i = 1, \ldots, N^* \\ \beta x_{(i)} + \tau\varepsilon_i & i = N^* + 1, \ldots, N, \end{cases} \tag{2.3}$$

or

$$y_{[i]} = \begin{cases} x_{(i)} + \tau\sqrt{x_{(i)}}\varepsilon_i & i = 1, \ldots, N^* \\ \beta x_{(i)} + \tau\sqrt{x_{(i)}}\varepsilon_i & i = N^* + 1, \ldots, N, \end{cases} \tag{2.4}$$

where $\varepsilon_i$ is generated from a normal distribution with mean zero and variance 1 and $y_{[i]}$ is the value of the $Y$-variable that corresponds to the value of $x_{(i)}.$

For a given integer $N,$ this model generates $N$ pairs of $(Y, X)$-measurements, for which the values of the $Y$-variable are proportional to the values of the $X$-variable. For the population units producing the largest $100(1-\omega)\,\%$ of the $Y$-values, the slope of the regression line between the $Y$- and $X$-variables is $\beta$ times larger than the slope of the regression line for other units. The variance of the $Y$-variable is constant in model (2.3) and increases with the $X$-values in model (2.4).

We performed a simulation study to investigate the efficiency of the PPS-ranked-set sample estimator. Finite populations of size $N = 2{,}000$ are generated from models (2.3) and (2.4). The slope parameter $\beta$ is selected to be 2 or 3. The parameter $\tau$ controls the correlation between the $Y$- and $X$-variables, $\rho = \text{cor}(X, Y),$ and is selected to be $\tau = 3, 8, 20.$ The parameter $\omega$ controls the percentage of population units having a larger proportionality constant for the units with the extreme $Y$-values. We consider $\omega$ values of 0.05, 0.10 and 0.20. For this population setting, we compare the efficiency of the PPS-ranked-set sample estimator with the PPS and ratio estimators of the population mean. The PPS with replacement samples were generated using the Lahiri (1951) method, which does not select any of the units with probability one from the population, and gives every unit in the population a positive probability of being selected in the sample. For each sample, the sample size is fixed at $n = dH,$ with

$d = 5$ and $H = 5, 10.$ Relatively smaller sample sizes $(n = 25,\ 50)$ are used to asses the small sample behaviors of the coverage probabilities of the confidence intervals of population mean. Simulation size is taken to be 20,000. Since the ratio estimator is not a design unbiased estimator, we use the mean squared error (MSE) of the ratio estimator to compare its efficiency with the PPS-ranked-set sample estimator. The MSE of the ratio estimator is computed as follows

$$\text{MSE}_R = \frac{1}{20{,}000} \sum_{i=1}^{20{,}000} \left(\bar{Y}_{R,\,i} - \mu_N\right)^2,\ \bar{Y}_{R,\,i} = \frac{\bar{Y}_i}{\bar{X}_i}\,\mu_X,$$

where $\bar{Y}_i$ and $\bar{X}_i$ are the sample means of the $Y$ - and $X$ -variables, respectively, in the $i^{\text{th}}$ iteration of the simulation, and $\mu_X$ is the population mean of the $X$ -variable.

**Table 2.2**
**Relative efficiency of the PPS-ranked-set sample estimator and coverage (COV) probability of the associated confidence interval for the population mean**

| Constant variance model eq. 2.3 | | | | | | | Increasing variance model, eq. 2.4 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\beta$ | $\omega$ | $\rho$ | $H$ | $\frac{\text{MSE}_R}{\sigma^2_{\bar{Y}_{\text{PR},N}}}$ | $\frac{\sigma^2_{\bar{Y}_{P,N}}}{\sigma^2_{\bar{Y}_{\text{PR},N}}}$ | COV | $\beta$ | $\omega$ | $\rho$ | $H$ | $\frac{\text{MSE}_R}{\sigma^2_{\bar{Y}_{\text{PR},N}}}$ | $\frac{\sigma^2_{\bar{Y}_{P,N}}}{\sigma^2_{\bar{Y}_{\text{PR},N}}}$ | COV |
| 2 | 0.05 | 0.933 | 5 | 5.385 | 1.657 | 0.939 | 2 | 0.05 | 0.918 | 5 | 3.335 | 1.348 | 0.949 |
| 2 | 0.05 | 0.933 | 10 | 7.281 | 2.231 | 0.936 | 2 | 0.05 | 0.918 | 10 | 4.012 | 1.587 | 0.950 |
| 2 | 0.05 | 0.932 | 5 | 4.338 | 1.523 | 0.945 | 2 | 0.05 | 0.844 | 5 | 1.607 | 1.090 | 0.950 |
| 2 | 0.05 | 0.932 | 10 | 5.389 | 1.907 | 0.944 | 2 | 0.05 | 0.844 | 10 | 1.731 | 1.144 | 0.950 |
| 2 | 0.05 | 0.926 | 5 | 2.093 | 1.235 | 0.952 | 2 | 0.05 | 0.611 | 5 | 1.128 | 1.018 | 0.947 |
| 2 | 0.05 | 0.926 | 10 | 2.158 | 1.358 | 0.951 | 2 | 0.05 | 0.611 | 10 | 1.167 | 1.035 | 0.949 |
| 2 | 0.10 | 0.951 | 5 | 5.136 | 1.900 | 0.930 | 2 | 0.10 | 0.939 | 5 | 3.404 | 1.533 | 0.952 |
| 2 | 0.10 | 0.951 | 10 | 6.982 | 2.567 | 0.938 | 2 | 0.10 | 0.939 | 10 | 4.017 | 1.778 | 0.949 |
| 2 | 0.10 | 0.951 | 5 | 4.283 | 1.748 | 0.941 | 2 | 0.10 | 0.875 | 5 | 1.660 | 1.157 | 0.952 |
| 2 | 0.10 | 0.951 | 10 | 5.339 | 2.188 | 0.945 | 2 | 0.10 | 0.875 | 10 | 1.718 | 1.181 | 0.949 |
| 2 | 0.10 | 0.946 | 5 | 2.227 | 1.372 | 0.953 | 2 | 0.10 | 0.648 | 5 | 1.134 | 1.040 | 0.950 |
| 2 | 0.10 | 0.946 | 10 | 2.273 | 1.490 | 0.950 | 2 | 0.10 | 0.648 | 10 | 1.129 | 1.029 | 0.948 |
| 2 | 0.20 | 0.975 | 5 | 4.005 | 1.989 | 0.939 | 2 | 0.20 | 0.965 | 5 | 2.888 | 1.631 | 0.948 |
| 2 | 0.20 | 0.975 | 10 | 5.253 | 2.764 | 0.941 | 2 | 0.20 | 0.965 | 10 | 3.316 | 1.941 | 0.952 |
| 2 | 0.20 | 0.974 | 5 | 3.419 | 1.843 | 0.942 | 2 | 0.20 | 0.911 | 5 | 1.581 | 1.210 | 0.950 |
| 2 | 0.20 | 0.974 | 10 | 4.100 | 2.370 | 0.947 | 2 | 0.20 | 0.911 | 10 | 1.598 | 1.238 | 0.949 |
| 2 | 0.20 | 0.970 | 5 | 1.873 | 1.442 | 0.950 | 2 | 0.20 | 0.711 | 5 | 1.151 | 1.067 | 0.951 |
| 2 | 0.20 | 0.970 | 10 | 1.819 | 1.585 | 0.954 | 2 | 0.20 | 0.711 | 10 | 1.123 | 1.047 | 0.949 |
| 3 | 0.05 | 0.873 | 5 | 5.560 | 1.679 | 0.936 | 3 | 0.05 | 0.867 | 5 | 4.700 | 1.551 | 0.945 |
| 3 | 0.05 | 0.873 | 10 | 7.596 | 2.286 | 0.935 | 3 | 0.05 | 0.867 | 10 | 6.130 | 1.998 | 0.945 |
| 3 | 0.05 | 0.873 | 5 | 5.220 | 1.636 | 0.940 | 3 | 0.05 | 0.832 | 5 | 2.697 | 1.253 | 0.950 |
| 3 | 0.05 | 0.873 | 10 | 6.973 | 2.178 | 0.938 | 3 | 0.05 | 0.832 | 10 | 3.121 | 1.414 | 0.950 |
| 3 | 0.05 | 0.870 | 5 | 3.862 | 1.462 | 0.947 | 3 | 0.05 | 0.687 | 5 | 1.416 | 1.062 | 0.949 |
| 3 | 0.05 | 0.870 | 10 | 4.610 | 1.774 | 0.946 | 3 | 0.05 | 0.687 | 10 | 1.504 | 1.100 | 0.950 |
| 3 | 0.10 | 0.914 | 5 | 5.274 | 1.924 | 0.926 | 3 | 0.10 | 0.909 | 5 | 4.601 | 1.786 | 0.944 |
| 3 | 0.10 | 0.914 | 10 | 7.257 | 2.632 | 0.937 | 3 | 0.10 | 0.909 | 10 | 5.969 | 2.289 | 0.945 |
| 3 | 0.10 | 0.914 | 5 | 5.005 | 1.877 | 0.934 | 3 | 0.10 | 0.880 | 5 | 2.791 | 1.402 | 0.953 |
| 3 | 0.10 | 0.914 | 10 | 6.717 | 2.505 | 0.940 | 3 | 0.10 | 0.880 | 10 | 3.144 | 1.551 | 0.950 |
| 3 | 0.10 | 0.912 | 5 | 3.875 | 1.674 | 0.945 | 3 | 0.10 | 0.747 | 5 | 1.451 | 1.111 | 0.951 |
| 3 | 0.10 | 0.912 | 10 | 4.635 | 2.027 | 0.946 | 3 | 0.10 | 0.747 | 10 | 1.479 | 1.119 | 0.949 |
| 3 | 0.20 | 0.957 | 5 | 4.090 | 2.009 | 0.936 | 3 | 0.20 | 0.953 | 5 | 3.694 | 1.886 | 0.944 |
| 3 | 0.20 | 0.957 | 10 | 5.434 | 2.825 | 0.940 | 3 | 0.20 | 0.953 | 10 | 4.664 | 2.497 | 0.948 |
| 3 | 0.20 | 0.957 | 5 | 3.919 | 1.968 | 0.940 | 3 | 0.20 | 0.929 | 5 | 2.446 | 1.490 | 0.950 |
| 3 | 0.20 | 0.957 | 10 | 5.073 | 2.703 | 0.942 | 3 | 0.20 | 0.929 | 10 | 2.680 | 1.680 | 0.952 |
| 3 | 0.20 | 0.956 | 5 | 3.125 | 1.768 | 0.946 | 3 | 0.20 | 0.815 | 5 | 1.414 | 1.155 | 0.950 |
| 3 | 0.20 | 0.956 | 10 | 3.588 | 2.194 | 0.949 | 3 | 0.20 | 0.815 | 10 | 1.409 | 1.162 | 0.949 |

Table 2.2 presents the efficiency results and the coverage (COV) probabilities of the approximate 95%-confidence intervals for the population mean based on the PPS-ranked-set sample mean. The efficiency results show that the PPS-ranked-set sample estimator has higher efficiencies than the PPS and ratio estimators for all simulation parameters in Table 2.2. The efficiency increases with each of the simulation parameters $\rho, H, \beta$ for fixed values of all the other parameters. For example, in the constant variance model, for fixed values of $\beta = 2,$ $\omega = 0.05$ and $H = 5,$ the efficiency values with respect to the ratio and PPS estimators are 2.093 and 1.235 for $\rho = 0.926$ and 5.385 and 1.657 for $\rho = 0.933,$ respectively. The same efficiency values in the increasing variance model are 1.128 and 1.018 for $\rho = 0.611$ and 3.335 and 1.348 for $\rho = 0.918,$ respectively. Similar observations can be made for other combination of simulation parameters.

The coverage probabilities of the confidence intervals for the population mean are relatively close to the nominal coverage probability 0.95 for both the constant and increasing variance models.

# 3 The PPS-ranked-set sample from stratified populations

In this section, we construct a PPS-ranked-set sample from a stratified population. The entire population is divided into $L$ stratum populations, $\mathcal{P}^{N_l} = \{u_{1,l}, \ldots, u_{N_l,l}\},$ where $N_l$ is the population size for the $l^{\text{th}}$ stratum population, $l = 1, \ldots, L.$ The stratum population means, variances and totals are given by

$$\mu_{N_l} = \frac{1}{N_l} \sum_{i=1}^{N_l} y_{i,l}, \;\; \sigma_{N_l}^2 = \frac{1}{N_l} \sum_{i=1}^{N_l} \left(y_{i,l} - \mu_{N_l}\right)^2, \;\; t_l = N_l \mu_{N_l}, \;\; l = 1, \ldots, L,$$

where $y_{i,l}$ is the value of $Y$ on unit $u_{i,l}$ in population $\mathcal{P}^{N_l}.$ The mean, total and variance of the overall population are defined as follows

$$\mu_N = \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_l} y_{i,l}, \;\; t_N = N\mu_N, \quad \sigma_N^2 = \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_l} \left(y_{i,l} - \mu_N\right)^2,$$

where $N = \sum_{l=1}^{L} N_l.$ The population total can be written as $t = N\mu_N.$ From this stratified population, we construct the stratified-PPS-ranked-set (SPR) sample

$$\left\{Y_{[h]i,l}, \pi_{[h]i,l}\right\}, \;\; i = 1, \ldots, d_l; \;\; h = 1, \ldots, H_l; \;\; l = 1, \ldots, L,$$

where $d_l$ and $H_l$ are the cycle and set sizes, respectively, in the stratum sample from population $l.$ Let $n_l = d_l H_l$ be the sample size for stratum $l, l = 1, \ldots, L.$ The estimators of the population mean and total are then given by

$$\bar{Y}_{\text{SPR},N} = \sum_{l=1}^{L} \frac{N_l}{N} \bar{Y}_{\text{PR},N_l} = \sum_{l=1}^{L} \frac{N_l}{N} \frac{1}{N_l H_l d_l} \sum_{h=1}^{H_l} \sum_{i=1}^{d_l} \frac{Y_{[h]i,l}}{\pi_{[h]i,l}}$$

and

$$T_{\text{SPR},N} = N\bar{Y}_{\text{SPR},N}.$$

If one ignores ranking information and uses PPS sampling, the stratified-PPS sample can be written as

$$\left\{Y_{i,\,l},\ \pi_{i,\,l}\right\},\quad i = 1, \ldots, n_l,\quad l = 1, \ldots, L.$$

The estimator of the population mean based on the stratified-PPS (SP) sample is given by

$$\bar{Y}_{\mathrm{SP},\,N} = \sum_{l=1}^{L} \frac{N_l}{N}\, \bar{Y}_{P,\,N_l} = \sum_{l=1}^{L} \frac{N_l}{N}\, \frac{1}{N_l n_l} \sum_{i=1}^{n_l} \frac{Y_{i,\,l}}{\pi_{i,\,l}}.$$

The variance of $\bar{Y}_{\mathrm{SP},\,N}$ can be found in standard text books

$$\sigma^2_{\bar{Y}_{\mathrm{SP},\,N}} = \sum_{l=1}^{L} \frac{N_l^2}{N^2}\, \frac{1}{n_l N_l^2} \sum_{k=1}^{N_l} \pi_k \left\{\frac{y_k}{\pi_k} - N_l \mu_{N_l}\right\}^2 = \sum_{l=1}^{L} \frac{N_l^2}{N^2}\, \sigma^2_{P,\,N_l}. \tag{3.1}$$

The next theorem shows that $\bar{Y}_{\mathrm{SPR},\,N}$ and $T_{\mathrm{SPR},\,N}$ are unbiased estimators for the population mean and total, respectively.

**Theorem 3.** *Let* $\left\{Y_{[h]i,\,l},\ \pi_{[h]i,\,l}\right\},\ i = 1, \ldots, d_l;\ h = 1, \ldots, H_l;\ l = 1, \ldots, L,$ *be a stratified-PPS-ranked-set sample from a stratified population. Under any consistent ranking model satisfying equation (2.1), the estimators* $\bar{Y}_{\mathrm{SPR},\,N}$ *and* $T_{\mathrm{SPR},\,N}$ *are unbiased for the population mean* $(\mu_N)$ *and total* $(t_N)$, *respectively, and their variances are given by* $\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}$ *and* $\sigma^2_{T_{\mathrm{SPR},\,N}}$,

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}} = \sum_{l=1}^{L} \frac{N_l^2}{N^2}\, \sigma^2_{\bar{Y}_{\mathrm{PR},\,N_l}} = \sum_{l=1}^{L} \frac{N_l^2}{N^2} \left\{\sigma^2_{P,\,N_l} - \frac{1}{N_l^2 n_l H_l} \sum_{h=1}^{H_l} \left(\mu_{[h:\,H_l]} - N_l \mu_{N_l}\right)^2\right\} \le \sigma^2_{\bar{Y}_{\mathrm{SP},\,N}}$$

$$\sigma^2_{T_{\mathrm{SPR},\,N}} = N^2 \sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}.$$

The proof of Theorem 3 follows from Theorem 1. Theorem 3 indicates that the variance of the sample mean $\bar{Y}_{\mathrm{SPR},\,N}$ based on a stratified-PPS-ranked-set sample is always less than or equal to the variance of the sample mean $\bar{Y}_{\mathrm{SP},\,N}$ based on the stratified-PPS-sample for settings where PPS sampling is appropriate.

From Theorem 2, an unbiased estimator for the variance $\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}$ is given by

$$\hat{\sigma}^2_{\bar{Y}_{\mathrm{SPR},\,N}} = \sum_{l=1}^{L} \frac{N_l^2}{N^2}\, \frac{1}{2 d_l^2 \left(d_l - 1\right) H_l^2 N_l^2} \sum_{h=1}^{H_l} \sum_{i=1}^{d_l} \sum_{j \ne i}^{d_l} \left\{\frac{Y_{[h]i,\,l}}{\pi_{[h]i,\,l}} - \frac{Y_{[h]j,\,l}}{\pi_{[h]j,\,l}}\right\}^2,\quad d_l > 1;\quad l = 1, \ldots, L.$$

This unbiased variance estimator provides a way to construct an approximate confidence interval for the population mean $\mu_N$, namely,

$$\bar{Y}_{\mathrm{SPR},\,N} \pm t_{df,\,\alpha/2}\, \hat{\sigma}_{\bar{Y}_{\mathrm{SPR},\,N}},$$

where $df = \sum_{l=1}^{L} n_l - \sum_{l=1}^{L} H_l$. For smaller sample sizes, the degrees of freedom can also be approximated using the Satterthwaite approximation to adjust for the effect of unequal stratum population variances. A similar expression can be written for a confidence interval for the population total $t_N$.

# 4 Sample size determination

One of the objectives of a stratified sampling design is to maximize the information content of the sample. Since our sampling design involves selecting samples from each one of the stratum populations, sample size allocation to strata populations becomes an important issue and has a big impact on the information content of the sample. We consider four different allocation methods: equal, proportional, Neyman and optimal allocation for a given cost model. The sample size allocation (as it relates to the efficiency of the estimator) depends very much on the cost structure of the sampling procedure and the magnitudes of the stratum-level variances. Hence, these four allocation procedures yield different efficiency results, since they try to minimize either the cost of sampling or the contribution of the stratum-level variances.

Note that the number of strata $L$ is fixed and the sample size for stratum population $l$ is $n_l$. For equal allocation, all stratum sample sizes are equal to $n_l \equiv n/L$, $l = 1, \ldots, L$, where $n$ is the total sample size in the stratified-PPS-ranked-set sample. For proportional allocation, the sample size $n_l$ is selected to be proportional to the stratum population size $N_l$ $l = 1, \ldots, L$, namely,

$$n_l = \frac{N_l}{N} n, \quad l = 1, \ldots, L.$$

Once $n_l$ is determined in this way, one can set $d_l = n_l / H_l$ for given set size $H_l$. For the setting where PPS sampling is appropriate, ranking in comparison sets is performed based on the size variable $X$. Since the variable $X$ is proportional to the variable $Y$, we expect that the $X$- and $Y$-variables are highly correlated. Hence, we select a moderately large value for $H_l$ for given $n_l$, such as $H_l = 5$, 6 or 10. Under proportional allocation, the variance of $\bar{Y}_{\mathrm{SPR},\,N}$ is given by

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(P) = \sum_{l=1}^{L} \frac{1}{NN_l n H_l} \sum_{h=1}^{H_l} \sigma^2_{[h:\,H_l]} = \sum_{l=1}^{L} \frac{\bar{\sigma}_l^2}{NN_l n}, \tag{4.1}$$

where $\bar{\sigma}_l^2 = \frac{1}{H_l} \sum_{h=1}^{H_l} \sigma^2_{[h:H_l]}$.

The Neyman allocation minimizes the variance of the estimator with respect to the sample size $n_l$ subject to the constraint that the sum of the stratum sample sizes equals $n$. Using Lagrange multipliers one can show that

$$n_l = \frac{n\bar{\sigma}_l}{\sum_{l=1}^{L} \bar{\sigma}_l}$$

minimizes the variance of $\bar{Y}_{\mathrm{SPR},\,N}$. Under a Neyman allocation, the variance of $\bar{Y}_{\mathrm{SPR},\,N}$ reduces to the simple form

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(N) = \sum_{l=1}^{L} \frac{\bar{\sigma}_l \sum_{l=1}^{L} \bar{\sigma}_l}{N^2 n}. \tag{4.2}$$

If the survey study has a budget constraint, the sample size allocation can be optimized by minimizing the variance of the estimator with a budgetary constraint in a cost function. A simple cost function for a stratified-PPS-ranked-set sample is given by

$$C_T = C_0 + \sum_{l=1}^{L} (c_l + r_l)\, n_l, \tag{4.3}$$

where $C_T$ is the total cost, $C_0$ is the overhead cost, $c_l$ is the cost of measuring a single observation from the stratum $l$ and $r_l$ is the cost of ranking $H_l$ observations in the comparison set in a stratum $l$. For settings where PPS-ranked-set sampling is appropriate, we expect that $r_l$ is either zero or very small. Under this cost function, the optimal allocation of the sample sizes is given by

$$n_l = n \frac{\bar{\sigma}_l \big/ \sqrt{c_l + r_l}}{\sum_{l=1}^{L} \bar{\sigma}_l \big/ \sqrt{c_l + r_l}}, \quad l = 1, \ldots, L.$$

Under the cost model (4.3), the variance of $\bar{Y}_{\mathrm{SPR},\,N}$ is given by

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(C) = \sum_{l=1}^{L} \frac{\bar{\sigma}_l \big/ \sqrt{c_l + r_l} \sum_{l=1}^{L} \bar{\sigma}_l \big/ \sqrt{c_l + r_l}}{N^2 n}.$$

We now compare the stratified-PPS-ranked-set sample estimator under the equal, proportional and Neyman allocation procedures. Under equal allocation, each stratum sample has the same sample size $n_l = n/L$, $l = 1, \ldots, L$. The variance of $\bar{Y}_{\mathrm{SPR},\,N}$ under the equal allocation is given by

$$\sigma^2_{\bar{X}_{\mathrm{SPR},\,N}}(E) = \sum_{l=1}^{L} \frac{L\bar{\sigma}_l^2}{N^2 n}. \tag{4.4}$$

The difference between the variances of $\bar{Y}_{\mathrm{SPR},\,N}$ under the equal and proportional allocations can be written as

$$\sigma^2_{\bar{Y}_{SPR,N}}(E) - \sigma^2_{\bar{Y}_{SPR,N}}(P) = \sum_{l=1}^{L} \frac{L\bar{\sigma}_l^2}{Nn} \left\{ \frac{N_l - \bar{N}}{NN_l} \right\}. \tag{4.5}$$

We expect that this difference would be positive for settings where large stratum populations have large variances. In that case, a proportional allocation increases the sample size for a large stratum population to reduce the contribution from this stratum sample to the variance of the estimator.

For Neyman allocation, we have

$$\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(E) - \sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(N) = \sum_{l=1}^{L} \frac{1}{N^2 n} \{\bar{\sigma}_l - \bar{\sigma}_.\}^2 \geq 0$$

and

$$\begin{aligned}
\sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(P) - \sigma^2_{\bar{Y}_{\mathrm{SPR},\,N}}(N) &= \sum_{l=1}^{L} \frac{\bar{\sigma}_l^2}{NN_l n} - \sum_{l=1}^{L} \frac{\bar{\sigma}_l \sum_{l=1}^{L} \bar{\sigma}_l}{N^2 n} \\
&= \sum_{l=1}^{L} \frac{L\bar{\sigma}_l^2}{\bar{N}N_l n} - \sum_{l=1}^{L} \frac{\bar{\sigma}_l \sum_{l=1}^{L} \bar{\sigma}_l}{N^2 n} \\
&\geq \sum_{l=1}^{L} \frac{L\bar{\sigma}_l^2}{N^2 n} - \sum_{l=1}^{L} \frac{\bar{\sigma}_l \sum_{l=1}^{L} \bar{\sigma}_l}{N^2 n} = \sum_{l=1}^{L} \frac{L}{N^2 n} \{\bar{\sigma}_l - \bar{\sigma}_.\}^2 \geq 0,
\end{aligned}$$

where $\bar{\sigma}_{\cdot}^{2} = \sum_{l=1}^{L} \bar{\sigma}_{l}^{2} / L.$ As expected, Neyman allocation always yields a smaller variance than both equal and proportional allocations, but it requires that the variance of the induced order statistics are known prior to construction of the sample. For the setting where the set sizes $H_l \equiv H$ for all stratum samples and the stratum population variances are known (or may be estimated) from previous studies, the Neyman allocation can be approximated as follows

$$n_l = \frac{n\bar{\sigma}_l}{\sum_{l=1}^{L} \bar{\sigma}_l} \approx \frac{n\hat{\sigma}_{N_l}}{\sum_{l=1}^{L} \hat{\sigma}_{N_l}}, \quad l = 1, \ldots, L,$$

where $\hat{\sigma}_{N_l}^{2}$ is the estimate of the variance, $\sigma_{N_l}^{2}$, of stratum population $l.$

# 5 Efficiency comparison of the new sampling design and estimator

In this section, we investigate the efficiency of the stratified-PPS-ranked-set sample estimators. We consider a stratified population with three strata $(L = 3).$ To see the effects of the stratum population sizes and variances on the allocation procedures, we generated stratified populations with different population sizes and variances. For clarity of notation, we define the proportions of population sizes and variances as follows:

$$p_{N_l} = \frac{N_l}{\sum_{l=1}^{L} N_l}, \quad p_{\sigma_{N_l}} = \frac{\sigma_{N_l}}{\sum_{l=1}^{L} \sigma_{N_l}}, \quad l = 1, \ldots, L.$$

We note that proportional and Neyman allocations select stratum sample sizes proportional to $p_{N_l}$ and $p_{\sigma_{N_l}},$ respectively.

In this part of the simulation, the population values of the $Y$- and $X$-variables are generated with a model different from the models in equations (2.3) and (2.4). For stratum population $l,$ we generate $\mathbf{X}_l^{*} = \left(X_{1,l}^{*}, \ldots, X_{N_l,l}^{*}\right)$ from

$$X_{i,l}^{*} = F^{-1}(i/(N+1); 0.1); \quad i = 1, \ldots, N_l,$$

where $F(;\lambda)$ is the cumulative distribution function of the exponential distribution with mean 10 (rate $\lambda = 0.1$). To simplify construction of the values of the $X$-variable from the stratum population, we re-scaled $X_{i,l}^{*}$ by

$$X_{i,l} = \sqrt{\frac{X_{i,l}^{*}}{\min(\mathbf{X}_l^{*})}}, \quad i = 1, \ldots, N_l.$$

The values of the variable $Y$ in the stratum population $l$ are generated from the quantiles of a normal distribution using the variable $X_{i,l}, i = 1, \ldots, N_l.$ We first compute

$$\varepsilon_{i,l} = G^{-1}(i/(N_l + 1); \theta_l, \tau_l); \quad i = 1, \ldots, N_l,$$

where $G(;a,b)$ is the CDF of a normal distribution with mean $a$ and standard deviation $b.$ The values of the $Y$-variable are then constructed from

$$Y_{i,\,l} \;=\; X_{i,\,l}\varepsilon_{i,\,l},\;\; i = 1, \ldots, N_l.$$

In this construction, it is clear that the values of the $Y$ -variable are proportional to the values of the $X$ -variable. Hence, use of the stratified-PPS-ranked-set sample would be appropriate.

For the simulation study, the total population size $N,$ the sample size $n$ and the location parameter $\theta_l\ (l = 1, 2, 3)$ are selected to be $N = 700,$ $n = 90,$ and $\theta_l = 5,$ respectively. The values $N_l$ and $\tau_l$ are varied to establish the values of $p_{N_l}$ and $p_{\sigma_{N_l}}$ in Tables 5.1 and 5.2. For the first four rows, the population sizes and standard deviations are selected to be $N_1 = 100,$ $N_2 = 200,$ $N_3 = 400$ and $\sigma_{N_1} = 35,$ $\sigma_{N_2} = 10,$ $\sigma_{N_3} = 5,$ respectively. For the last four rows, the population sizes and standard deviations are selected as $N_1 = 400,$ $N_2 = 200,$ $N_3 = 100$ and $\sigma_{N_1} = 35,$ $\sigma_{N_2} = 10,$ $\sigma_{N_3} = 5,$ respectively. To make the comparison easier, the same set size $H$ $(H = 2, 3, 5, 6)$ is used in all stratum populations for any combination of particular choices of $p_{N_l}$ and $p_{\sigma_{N_l}},$ $l = 1, \ldots, L.$

An unbiased estimator of the variance of the sample mean requires that $d_l \geq 2,$ for $l = 1, \ldots, L.$ In Neyman and proportional allocations, this assumption may not hold in certain stratum samples when $p_{N_l}$ or $p_{\sigma_{N_l}}$ is too small. In this case, we modified the Neyman and proportional allocations to make sure that $d_l \geq 2$ by reducing the maximum $d_l$ and increasing any $d_l$ smaller than 2. These allocation procedures may not be optimal under this modification.

**Table 5.1**
**Relative efficiencies of the stratified-PPS-sample (SP) with respect to the stratified-PPS-ranked-set (SPR) sample; E: Equal allocation; P: Proportional allocation; N: Neyman allocation**

| | **Proportion of $N_l$** | | | **Proportion of $\sigma^2_{N_l}$** | | | **Efficiencies** | | |
|---|---|---|---|---|---|---|---|---|---|
| $H$ | $p_{N_1}$ | $p_{N_2}$ | $p_{N_3}$ | $p_{\sigma^2_{N_1}}$ | $p_{\sigma^2_{N_2}}$ | $p_{\sigma^2_{N_3}}$ | $\dfrac{\sigma^2_{\bar{Y}_{\text{SP}}}(E)}{\sigma^2_{\bar{Y}_{\text{SPR}}}(E)}$ | $\dfrac{\sigma^2_{\bar{Y}_{\text{SP}}}(P)}{\sigma^2_{\bar{Y}_{\text{SPR}}}(P)}$ | $\dfrac{\sigma^2_{\bar{Y}_{\text{SP}}}(N)}{\sigma^2_{\bar{Y}_{\text{SPR}}}(N)}$ |
| 2 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.472 | 1.408 | 2.007 |
| 3 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.927 | 1.850 | 2.627 |
| 5 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 2.803 | 3.001 | 3.823 |
| 6 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 3.229 | 3.059 | 4.402 |
| 2 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 1.468 | 1.496 | 1.506 |
| 3 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 1.915 | 1.917 | 1.965 |
| 5 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 2.769 | 2.715 | 2.689 |
| 6 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 3.180 | 3.358 | 2.440 |

Table 5.1 presents the relative efficiencies of the stratified-PPS-ranked-set sample mean with respect to the stratified-PPS sample mean for the equal, proportional and Neyman allocation procedures. The efficiencies are computed using equations (3.1), (4.1), (4.2) and (4.4). It is clear that the stratified-PPS-ranked-set sample mean has higher efficiency than the stratified-PPS sample mean for all allocation procedures. The efficiency improvement increases with the set size $H.$

**Table 5.2**
**Relative efficiencies of the stratified-PPS-ranked-set sample estimator with respect to Neyman allocation and the coverage probabilities of confidence intervals; E: Equal allocation; P: Proportional allocation; N: Neyman allocation**

| | **Proportion of $N_l$** | | | **Proportion of $\sigma^2_{N_l}$** | | | **Efficiencies** | | **Coverage Prob** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $H$ | $p_{N_1}$ | $p_{N_2}$ | $p_{N_3}$ | $p_{\sigma^2_{N_1}}$ | $p_{\sigma^2_{N_2}}$ | $p_{\sigma^2_{N_3}}$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(E)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(N)}$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(P)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(N)}$ | **Eq** | **Prop** | **Neyman** |
| 2 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.021 | 1.358 | 0.951 | 0.947 | 0.946 |
| 3 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.021 | 1.354 | 0.950 | 0.945 | 0.948 |
| 5 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.021 | 1.214 | 0.949 | 0.950 | 0.953 |
| 6 | 0.143 | 0.286 | 0.571 | 0.726 | 0.161 | 0.113 | 1.021 | 1.372 | 0.950 | 0.933 | 0.948 |
| 2 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 2.327 | 1.357 | 0.941 | 0.947 | 0.945 |
| 3 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 2.325 | 1.381 | 0.944 | 0.949 | 0.951 |
| 5 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 2.201 | 1.334 | 0.939 | 0.946 | 0.949 |
| 6 | 0.571 | 0.286 | 0.143 | 0.945 | 0.047 | 0.008 | 1.739 | 0.979 | 0.941 | 0.943 | 0.944 |

Table 5.2 presents the efficiencies of the allocation procedures and the coverage probabilities of the approximate confidence interval for the population mean constructed from the stratified-PPS-ranked-set samples. Again the efficiencies are computed from the analytic expressions in equations (3.1), (4.1), (4.2) and (4.4), but the coverage probabilities are computed from a simulation study by generating 5,000 stratified-PPS-rankek-set samples. The PPS samples are generated using the function 'lahiri.design' in the R-package SDaA, Verbeke (2014). Efficiencies of the equal and proportional allocations are compared with respect to the Neyman allocation. Since the Neyman allocation is optimal, we see that all entries, except 0.979 in the last row of column 9, are greater than 1, as expected. The reason that the proportional allocation is better than the Neyman allocation in the last row is that the Neyman allocation is modified. The Neyman allocation yields $d_1 = 14$, $d_2 = 1$, $d_3 = 0$. This allocation is modified to $d_1 = 11$, $d_2 = 2$, $d_3 = 2$ so that the cycle size in each stratum sample is greater than 1. The proportional allocation in the last row did not need any modification. Since the Neyman allocation is no longer optimal in this case, it is not as efficient as the proportional allocation.

Neyman allocation is always better than equal allocation even when we modify it for the cycle sizes. The efficiency of proportional allocation with respect to equal allocation can be obtained by dividing column 8 by column 9 in Table 5.2. If the ratio of the entries in column 8 and column 9 is greater than 1, proportional allocation is more efficient than equal allocation.

It is clear that in the first 4 rows of Table 5.2, equal allocation is better than proportional allocation. In these populations, smaller stratum populations have larger variances. Hence, proportional allocation selects less data from the stratum having large variance and more data from the stratum having small variance. In the last four rows of Table 5.2, where large populations have large variances, proportional allocation has higher efficiency than equal allocation since it allocates larger sample sizes to strata with larger variances. These are consistent with the finding in equation (4.5), which indicates that proportional allocation is more efficient when large stratum populations have large variances.

The last three columns of Table 5.2 provide the coverage probabilities of the confidence intervals for the population mean for equal, proportional and Neyman allocation procedures. It is clear that all coverage probabilities are very close to the nominal coverage probability 0.95.

# 6 Example

In this section, we apply the stratified PPS-ranked-set sample design to apple production data. We considered that the apple production data provided by Turkish Statistical Institute is a finite population. The apple farms in this population are divided into seven different $(L = 7)$ geographical regions. The farms in each region are considered a stratum population. Table 6.1 indicates that the population sizes $(N_l)$, means $(\mu_{N_l})$ and standard deviations $(\sigma_{N_l})$ vary significantly. Hence, it is natural to use stratified sampling to reduce the sampling variation. Since the number of apple trees $(X)$ in each township is available and the correlation coefficients $(\rho_l)$ between the $X$ - and $Y$ -variables are relatively high in the stratum populations, the PPS-ranked-set sample in each stratum population would be a reasonable choice.

**Table 6.1**
**Population characteristics of the apple production data (in tons, 1 ton = 1,000kg)**

| Strata $(l)$ | $\mu_l$ | $\sigma_l$ | $N_l$ | $\rho_l$ |
|---|---|---|---|---|
| Marmara $(l = 1)$ | 1,536.8 | 6,425 | 106 | 0.816 |
| Aegean $(l = 2)$ | 2,233.7 | 11,604.9 | 105 | 0.856 |
| Mediterranean $(l = 3)$ | 9,384.31 | 29,907.5 | 94 | 0.901 |
| Black Sea $(l = 4)$ | 967 | 2,389.7 | 204 | 0.713 |
| Central Anatolia $(l = 5)$ | 5,588 | 28,643.4 | 171 | 0.986 |
| Eastern Anatolia $(l = 6)$ | 631.4 | 1,171.1 | 103 | 0.885 |
| Southeastern Anatolia $(l = 7)$ | 72.4 | 111.3 | 68 | 0.917 |

We treated the apple production data as a stratified population and simulated stratified-PPS-ranked-set (SPR) samples for sample size $n = 210$ and set sizes $H = 2, 3, 5, 6.$ In order to compare the SPR with the competitor sampling designs, we also generated samples using stratified simple random sample (SSRS) and stratified-PPS (SP) sample. Samples are selected with the equal (E), proportional (P) and Neyman (N) allocation procedures in all three sampling designs.

For Neyman and proportional allocations, whenever the stratum cycle size $d_l$ is less than 2, the Neyman and proportional allocations are modified by changing all $d_l < 2$ to 2 and reducing the maximum $d_l$ so that the total sample size still equals $n$. This modification allows us to obtain an unbiased estimator for the variance of the stratified-PPS-ranked-set sample mean.

**Table 6.2**
**The relative efficiencies of the stratified SRS, PPS and PPS-ranked-set sample designs and the coverage probabilities of the approximate confidence interval for the population mean of the apple production data; E: Equal allocation; P: Proportional allocation; N: Neyman allocation**

| | **Stratified SRS** | | | **Stratified PPS** | | | **Stratified PPS-RSS** | | **Coverage Prob** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $H$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SSRS}}}(E)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(E)}$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SSRS}}}(P)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(P)}$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SSRS}}}(N)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(N)}$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SP}}}(E)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(E)}$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SP}}}(P)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(P)}$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SP}}}(N)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(N)}$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(E)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(N)}$ | $\frac{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(P)}{\sigma^2_{\bar{\bar{Y}}_{\text{SPR}}}(N)}$ | **Equ.** | **Prop.** | **Neyman** |
| 2 | 1.139 | 1.154 | 1.321 | 1.143 | 1.158 | 1.259 | 1.777 | 1.901 | 0.943 | 0.939 | 0.948 |
| 3 | 1.254 | 1.251 | 1.447 | 1.243 | 1.243 | 1.351 | 1.759 | 1.872 | 0.945 | 0.943 | 0.952 |
| 5 | 1.399 | 1.371 | 1.464 | 1.399 | 1.470 | 1.326 | 1.566 | 1.668 | 0.947 | 0.942 | 0.949 |
| 6 | 1.475 | 1.454 | 1.434 | 1.437 | 1.435 | 1.294 | 1.434 | 1.580 | 0.944 | 0.938 | 0.947 |

Table 6.2 provides the relative efficiencies of the PPS-ranked-set sample mean with respect to other competing estimators. These entries are obtained from a simulation of 10,000 replications. It is clear that the SPR sample mean is more efficient than both SSRS and SP sample means for all allocation procedures. Efficiencies in general increase with the set size $H$ when the stratum cycle sizes $(d_l = n_l / H > 1)$ are not too small ($H$ is not large). When $H$ is large, the Neyman allocation is modified to make sure that $d_l \geq 2$. In this case, the modified Neyman allocation loses its optimality properties, but it is still better than the other allocation procedures.

In Section 4, we observed that if the large populations have large variances, proportional allocation is better than equal allocation. Table 6.1 indicates that some of the smaller stratum populations have very large variances. For example, the population in the Mediterranean region (the second smallest population) has 94 farms, but its standard deviation is the largest among the 7 stratum populations. Hence, the efficiencies of the proportional and equal allocations in the SSRS and SP samples appear to be the same with respect to the SPR samples (columns 2, 3 for SSRS and columns 5, 6 for SP).

For the stratified-PPS-ranked-set sample, Neyman allocation provides a substantial amount of improvement over the equal and proportional allocations. The efficiency improvement is a deceasing function of the set size $H$, but this reduction is due to the use of the modified Neyman allocation for large set sizes (i.e, small $d_l$).

In the stratified-PPS-ranked-set sample, equal allocation is more efficient than proportional allocation. This can be seen from the ratio of columns 8 and 9. If we divide column 8 by column 9, all entries would be less than 1, which indicates that equal allocation has smaller variance than proportional allocation. This is again consistent with equation (4.5), which shows that equal allocation is better if the large stratum populations have small variances.

Table 6.2 also provides the coverage probabilities of approximate 95% confidence intervals for the population mean. It is clear that the coverage probabilities for all allocation procedures are very close to the nominal coverage probability of 0.95.

# 7 Concluding remarks

A probability-proportional-to-size sampling provides highly efficient estimators for the population mean and total when there exists a size measure for every unit in the population. The size measure contains significant information about the importance of each unit being included in the sample. It also provides important information about the relative position (rank) of the population units. Combining these two pieces of information in a meaningful way leads to a new sampling design, the stratified-PPS-ranked-set sampling. The stratified-PPS-ranked-set sampling combines the efficiency gains of the probability sampling and the position (rank) information of the sample unit in a comparison set.

We constructed unbiased estimators for the population mean, total and their variances. The sample size allocation to each stratum plays a significance role in the efficiency of the estimators. The choice of the sample size allocation depends on sampling cost, stratum population sizes and variances. If the larger populations have larger variances, proportional allocation works reasonably well. The new sampling design is applied to an apple production data in a stratified population.

# Appendix

**Proof of Theorem 1:** We first note that $\left(Y_{[h]i},\, \pi_{[h]i}\right),\, i = 1, \ldots, d,$ are iid random variables. We then write

$$\begin{aligned} E\left(\bar{Y}_{\mathrm{PR},\,N}\right) &= \frac{1}{dHN} \sum_{h=1}^{H} dE\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right) = \frac{1}{dHN} \sum_{h=1}^{H} \sum_{k=1}^{N} dy_k \, \frac{P\left(Y_{[h]1} = y_k\right)}{\pi_k} \\ &= \frac{1}{HN} \sum_{h=1}^{H} \sum_{k=1}^{N} \frac{y_k}{\pi_k} \, f_{[h:H]}\left(y_k\right) = \frac{1}{HN} \sum_{k=1}^{N} \frac{y_k}{\pi_k} \sum_{h=1}^{H} f_{[h:H]}\left(y_k\right). \end{aligned}$$

Using the consistency of the ranking procedure in equation (2.1), we write

$$E\left(\bar{Y}_{\mathrm{PR},\,N}\right) = \frac{1}{HN} \sum_{k=1}^{N} \left(\frac{y_k}{\pi_k}\right) Hf\left(y_k\right) = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{y_k}{\pi_k}\right) \sum_{j=1}^{N} \pi_j I\left(y_k = y_j\right) = \frac{1}{N} \sum_{k=1}^{N} y_k = \mu_N .$$

For the proof of the variance, consider

$$\begin{aligned} \operatorname{Var}\left(\bar{Y}_{\mathrm{PR},\,N}\right) &= \frac{d}{H^2 d^2 N^2} \sum_{h=1}^{H} \operatorname{Var}\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right) \\ &= \frac{1}{H^2 d N^2} \sum_{h=1}^{H} \left\{ E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \left[E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)\right]^2 \right\} \\ &= \frac{1}{nHN^2} \sum_{h=1}^{H} E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \frac{1}{nHN^2} \sum_{h=1}^{H} \mu_{[h:H]}^2 . \end{aligned} \tag{A.1}$$

Again using the consistency of the within-set ranking procedure in equation (2.1), we write

$$\begin{aligned} \frac{1}{nN^2 H} \sum_{h=1}^{H} E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 &= \frac{1}{nN^2} E\left(\frac{Y_1}{\pi_1}\right)^2 = \frac{1}{nN^2} \operatorname{Var}\left(\frac{Y_1}{\pi_1}\right) + \frac{1}{nN^2} \left\{E\left(\frac{Y_1}{\pi_1}\right)\right\}^2 \\ &= \operatorname{Var}\left(\bar{Y}_{\mathrm{PPS}}\right) + \mu_N^2 / n . \end{aligned}$$

We now insert this result in equation (A.1) and write

$$\begin{aligned}\operatorname{Var}\left(\bar{Y}_{\mathrm{PR},\,N}\right) &= \operatorname{Var}\left(\bar{Y}_{\mathrm{PPS}}\right) + \mu_N^2/n - \frac{1}{nHN^2}\sum_{h=1}^{H}\mu_{[h:H]}^2 \\ &= \operatorname{Var}\left(\bar{Y}_{\mathrm{PPS}}\right) - \frac{1}{nHN^2}\sum_{h=1}^{H}\left(\mu_{[h:H]} - N\mu_N\right)^2 \\ &\leq \operatorname{Var}\left(\bar{Y}_{\mathrm{PPS}}\right).\end{aligned}$$

This completes the proof.

**Proof of Theorem 2:** It is easy to see that

$$\begin{aligned}E\left(\hat{\sigma}^2_{\bar{Y}_{\mathrm{PR},\,N}}\right) &= \frac{2d\,(d-1)}{2d^2\,(d-1)\,N^2H^2}\sum_{h=1}^{H}E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \frac{2d\,(d-1)}{2d^2\,(d-1)\,N^2H^2}\sum_{h=1}^{H}E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)E\left(\frac{Y_{[h]2}}{\pi_{[h]2}}\right) \\ &= \frac{1}{dH^2N^2}\sum_{h=1}^{H}\left\{E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)^2 - \left[E\left(\frac{Y_{[h]1}}{\pi_{[h]1}}\right)\right]^2\right\} = \frac{1}{dH^2N^2}\sum_{h=1}^{H}\sigma^2_{[h:H]}.\end{aligned}$$

# References

Al-Saleh, M.F., and Samawi, H.M. (2007). A note on inclusion probability in ranked set sampling and some of its variations. *Test*, 16, 198-209.

Deshpande, J.V., Frey, J. and Ozturk, O. (2006). Nonparametric ranked-set sampling confidence intervals for a finite population. *Environmental and Ecological Statistics*, 13, 25-40.

Frey, J. (2011). Recursive computation of inclusion probabilities in ranked set sampling. *Journal of Statistical Planning and Inference*, 141, 3632-3639.

Gokpinar, F., and Ozdemir, Y.A. (2010). Generalization of inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 39, 89-95.

Kadilar, C., and Cingi, H. (2003). Ratio estimators in stratified random sampling. *Biometrical Journal*, 45, 218-225.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute*, 33, 133-140.

MacEachern, S.N., Stasny, E.A. and Wolfe, D.A. (2004). Judgment post-stratification with imprecise rankings. *Biometrics*, 60, 207-215.

Muttlak, H.A., and McDonald, L.L. (1992). Ranked set sampling and the line intercept method: A more efficient procedure. *Biometrical Journal*, 34, 329-346.

Nematollahi, N., Salehi, M.M. and Aliakbari Saba, R. (2008). Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. *Communications in Statistics - Theory and Methods*, 37, 2402-2415.

Ozdemir, Y.A., and Gokpinar, F. (2007). A generalized formula for inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 36, 89-99.

Ozdemir, Y.A., and Gokpinar, F. (2008). A new formula for inclusion probabilities in median ranked set sampling. *Communications in Statistics - Theory and Methods*, 37, 2022-2033.

Ozturk, O. (2014). Estimation of population mean and total in a finite population setting using multiple auxiliary variables. *Journal of Agricultural, Biological, and Environmental Statistics*, 19, 161-184.

Ozturk, O. (2016). Statistical inference based on judgment post-stratified samples in finite population. *Survey Methodology*, 42, 2, 239-262. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016002/article/14664-eng.pdf.

Ozturk, O. (2019a). Two-stage cluster samples with ranked set sampling designs. *Annals of the Institute of Statistical Mathematics*, 71, 63-91.

Ozturk, O. (2019b). Post-stratified probability-proportional-to-size sampling from stratified populations. *Journal of Agricultural, Biological, and Environmental Statistics*, https://doi.org/10.1007/s13253-019-00370-6.

Ozturk, O., and Bayramoglu-Kavlak, K. (2018). Statistical inference using stratified ranked set samples from finite populations. *Ranked Set Sampling: 65 Years Improving the Accuracy in Data Gathering*, (Eds., C. Bouza and A.I. Al-Omari), 157-170, Elsevier, New York.

Ozturk, O., and Jafari Jozani, M. (2013). Inclusion probabilities in partially rank ordered set sampling. *Computational Statistics and Data Analysis*, 69, 122-132.

Patil, G.P., Sinha, A.K. and Taillie, C. (1995). Finite population corrections for ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 47, 621-36.

Presnell, B., and Bohn, L.L. (1999). U-Statistics and imperfect ranking in ranked set sampling. *Journal of Nonparametric Statistics*, 10, 111-126.

Sroka, C.J. (2008). Extending ranked set sampling to survey methodology. PhD thesis, Department of Statistics, Ohio State University.

Sud, V., and Mishra, D.C. (2006). Estimation of finite population mean using ranked set two stage sampling designs. *Journal of the Indian Society of Agricultural Statistics*, 60, 108-117.

Thompson, S.K. (2002). *Sampling*, New York: John Wiley & Sons, Inc.

Verbeke, T. (2014). SDaA: Sampling: Design and Analysis. R package version 0.1-3, https://CRAN.R-project.org/package=SDaA.

Wang, X., Lim, J. and Stokes, L. (2016). Using ranked set sampling with cluster randomized designs for improved inference on treatment effects. *Journal of the American Statistical Association*, 516, 1576-1590.

Wolfe, D. (2012). Ranked set sampling: Its relevance and impact on statistical inference. *International Scholarly Research Network ISRN Probability and Statistics*, ID 568385, doi:10.5402/2012/568385.

# Semi-automated classification for multi-label open-ended questions

**Hyukjun Gweon, Matthias Schonlau and Marika Wenemark[1]**

## Abstract

In surveys, text answers from open-ended questions are important because they allow respondents to provide more information without constraints. When classifying open-ended questions automatically using supervised learning, often the accuracy is not high enough. Alternatively, a semi-automated classification strategy can be considered: answers in the easy-to-classify group are classified automatically, answers in the hard-to-classify group are classified manually. This paper presents a semi-automated classification method for multi-label open-ended questions where text answers may be associated with multiple classes simultaneously. The proposed method effectively combines multiple probabilistic classifier chains while avoiding prohibitive computational costs. The performance evaluation on three different data sets demonstrates the effectiveness of the proposed method.

**Key Words:** Semi-automated classification; Open-ended questions; Multi-label data.

# 1 Introduction

Open-ended questions in surveys are often manually classified into different class or categories. When data are large, manual classification is time consuming and expensive in the sense that it requires professional human coders with sufficient knowledge. At the same time, analyzing the text answers from open-ended questions is important because they do not constrain respondents' answers and thus may give more accurate information than closed-ended questions (Schonlau and Couper, 2016).

The advance of statistical learning techniques can be used for automatic classification for text data from open-ended questions. A statistical learning model such as Support Vector Machines (SVM) (Vapnik, 2000) and Random Forests (Breiman, 2001) may be trained based on training data and used to predict new data. Analyzing text data from open-ended questions with statistical learning methods has received increasing attention in social sciences (Matthews, Kyriakopoulos and Holcekova, 2018; Ye, Medway and Kelley, 2018).

While the use of statistical learning methods reduces the total cost for the coding task, fully automated classification for open-ended questions remains challenging. It is often difficult to achieve an overall classification accuracy as high as the accuracy that can be achieved by human coders and with a classification accuracy which is acceptable to use for research purposes. Semi-automated classification uses statistical approaches to partially automated classification in that easy-to-classify answers are categorized automatically and hard-to-classify answers are categorized manually. (Gweon, Schonlau, Kaczmirek, Blohm and Steiner, 2017; Schonlau and Couper, 2016).

Answers to open-ended questions are often associated with multiple categories simultaneously. In the community of machine learning, this type of data is referred to as multi-label data. This is different from

[1] Hyukjun Gweon, Department of Statistical and Actuarial Sciences, Western University, London, Ontario, N6A 5B7, Canada. E-mail: hgweon@uwo.ca; Matthias Schonlau, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada; Marika Wenemark, Centre for Organisational Support and Development, Linkping University, Region stergtland, Sweden.

the traditional multi-class data where a text answer can only belong to a single class or label. Recently, Schonlau, Gweon and Wenemark (2019) evaluated the use of existing machine learning algorithms for fully automated coding of multi-label open-ended questions.

This paper focuses on semi-automated classification for multi-labelled text data from open-ended questions. As far as we are aware, there is no published work on semi-automated classification for multi-label data. Most of the previous work on semi-automated classification deal with multi-class data. Also most research in machine learning that analyzes multi-label data assumes fully automated classification. In this paper we consider existing algorithms for multi-label data that may be suitable for semi-automatic classification. We also propose a new method to improve the classification performance of existing methods in the specific context of multi-label semi-automatic classification. This is illustrated with three examples of multi-labelled text data from open-ended questions. We show that the proposed method can achieve a higher accuracy than Binary Relevance, Label Powerset, and Probabilistic Classifier Chains (Dembczyński, Cheng and Hüllermeier, 2010) for semi-automated classification.

The rest of this paper is organized as follows: In Section 2, we review elements of semi-automated classification for open-ended questions. In Section 3, we review approaches to multi-label classification. In Section 4, we present the details of the proposed approach. In Section 5, we evaluate the proposed method as well as other commonly used algorithms based on multi-label text data from open-ended questions. In Section 6, we conclude with a discussion.

# 2 Semi-automated classification for text data

This section describes how text answers to open ended-questions are converted into ngram variables and how a learning algorithm is evaluated in semi-automated classification.

## 2.1 Converting text answers into ngram variables

To use text answers as the input features for a learning algorithm, we may transform the original texts into a different representation using text mining approaches. A common transformation approach is to create indicator variables, each of which indicates the presence or absence of a certain word (unigram) or a short word sequence (bigram, or more generally, ngram variables) (Sebastiani, 2002; Schonlau, Guenther and Sucholutsky, 2017). Applying this technique, we may convert any text answer into a vector in which each element is binary and corresponds to a word (or a word sequence). Instead of indicator variables, variables containing word frequency can also be used (Manning, Raghavan and Schütze, 2008; Guenther and Schonlau, 2016).

Typically, there are several thousands of ngram variables including redundant words. We may reduce the number of ngram variables by applying some preprocessing techniques such as stemming (i.e., reducing words to their grammatical root) and thresholding (i.e., removing words occurred less than a certain time) and removing very common words (stopwords) (Manning et al., 2008; Guenther and Schonlau, 2016).

## 2.2 Production rate

Semi-automated classification requires a score or a probability that shows a level of confidence about the prediction. A threshold on that score or probability divides the text answers into easy-to-classify and hard-to-classify texts. All new text answers with high scores above a threshold may be categorized automatically and all others are categorized manually. The threshold is a user-specified value and can be set depending on the combination of desired prediction accuracy in the easy-to-classify group and the acceptable number of difficult-to-classify answers that need manual coding. The production rate refers to as the fraction of text answers that belong to the easy-to-classify group. That is, the production rate is the proportion of observations that can be categorized automatically. In general, production rate and accuracy are inversely related. If we chose a low production rate, only the easiest answers will be in the easy-to-classify group and the accuracy of the automatic classification will be high. If we increase the production rate, more complicated answers will be automatically classified and accuracy will tend to decrease.

For multi-label data, the definition of accuracy is no longer obvious. Evaluation measures for multi-label data are discussed in Section 3.1.

# 3 Multi-label classification

Consider a set of possible output labels $\mathcal{L} = \{1, 2, \ldots, L\}$. In multi-label classification, each instance with a feature vector $\mathbf{x} \in \mathbb{R}^d$ is associated with a subset of these labels. Equivalently, the subset can be described as $\mathbf{Y} = (y_1, y_2, \ldots, y_L)$, where $y_i = 1$ if label $i$ is associated with the instance, and $y_i = 0$ otherwise. A multi-label classifier $\mathbf{h}$ learns from training data to predict $\mathbf{h}(\mathbf{x}) = \hat{\mathbf{Y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L)$ for a given $\mathbf{x}$.

Next, we review some common multi-label algorithms and their relationship to an evaluation criterion, subset accuracy.

## 3.1 Evaluating multi-label algorithms in semi-automated classification

Evaluating the classification of a text answer into a single label is straightforward: the label is either correct or not and accuracy refers to the percentage of correctly classified answers; equivalently, error refers to the percentage of misclassified answers. For answers that are classified into multiple labels, there are several ways to combine the accuracy of each single label to an overall evaluation measure for the set of multiple labels. These evaluation measures include subset accuracy, Hamming loss, F-measure and log loss. For a predicted set of multiple labels, subset accuracy is 1 if all of the $L$ labels are correctly predicted and 0 otherwise. Hamming loss evaluates the fraction of misclassified labels. F-measure is the harmonic mean of precision and recall and log loss evaluates the uncertainty of the prediction averaged over the labels when a probability score for each label is given.

In this paper we develop a methodology for subset accuracy (equivalently, in terms of loss, 0/1 loss). This is a strict metric because a zero score is given even if all labels are correctly classified except one. However, subset accuracy is appropriate for semi-automated classification because if an algorithm has difficulty classifying even a single label, the entire observation needs to be manually classified. That is, automated classification shall be conducted only if the model is highly confident in the entire predicted label set.

Because subset accuracy requires that all labels are simultaneously correctly classified, we are interested in finding the label set $Y^*$ that maximizes the joint probability conditional on a text answer $\mathbf{x}$:

$$Y^* = \text{argmax}_{\mathbf{Y}} P(\mathbf{Y} \mid \mathbf{x}) = \text{argmax}_{\mathbf{Y}} P(y_1, \ldots, y_L \mid \mathbf{x}).$$

In the next section we discuss common approaches to estimating the joint probability proposed in the machine learning community.

## 3.2 Multi-label approaches that optimize subset accuracy

Various approaches have been proposed for predicting multi-label outcomes. Since we use subset accuracy as the evaluation measure, we focus on methods that aim to maximize the joint conditional distribution.

The simplest approach, called Binary Relevance (BR), transforms a multi-label problem into separate binary problems. That is, BR constructs a binary classification model for each label independently. For an unseen observation, the prediction set of labels is obtained simply by combining the individual binary results. In other words, the predicted label set is the union of the results predicted from the $L$ binary models. If each of the binary models produces probability outcomes, BR can produce an estimate for $P(y_1 \mid \mathbf{x}) P(y_2 \mid \mathbf{x}) \ldots P(y_L \mid \mathbf{x})$. Note that this coincides with the joint probability $P(y_1, \ldots, y_L \mid \mathbf{x})$ if the labels are independent (conditional on $\mathbf{x}$). This implies that the product of the probabilities obtained by BR will estimate $P(y_1, \ldots, y_L \mid \mathbf{x})$ accurately only if the labels are conditionally independent. The joint probability may be inaccurate if the labels are substantially correlated given $\mathbf{x}$.

Another approach tailored for subset accuracy is Label Powerset learning (LP). This approach transforms a multi-label classification into a multi-class (i.e., multinomial) problem by treating each unique label set $\mathbf{Y}$ that exists in the training data as a single class. For example, when $L = 3$ there could be up to $2^3$ classes $c_i$, $(i = 1, \ldots, 8)$ observed in the training data. Then any algorithm for multi-class problems can be applied using the transformed $c_i$ classes. Training a multi-class classifier takes into consideration dependencies between labels. For a new observation, LP predicts the most probable class (i.e., the most probable label set). If an algorithm for multi-class data gives probabilistic outputs (some algorithms classify without computing probabilities), LP directly estimates the class probabilities (i.e., the joint probability $P(\mathbf{Y}\,|\,\mathbf{x})$). However, this approach cannot estimate the joint probability for any label set unseen in the training data. As a consequence, if the true label set of the new observation is an *unseen* observation the prediction cannot be correct. Another drawback of LP is that the number of classes in the transformed problem can increase exponentially (up to $2^L$ number of classes). This can be problematic when L is large since each combination of labels may be present in just one or a few observations in the training data which makes the learning process difficult.

A third approach to multi-label learning is Classifier Chains (CC) (Read, Pfahringer, Holmes and Frank, 2009, 2011). As in binary relevance, in CC also a binary model is fit for each label. However, CC fits the binary models sequentially and uses the binary label results obtained from previous models as additional predictors in subsequent models. That is, the model for the $i^{\text{th}}$ label $y_i$ uses $\mathbf{x}$ and $y_1, \ldots, y_{i-1}$ as features. (For example, the model for $y_1$ uses $\mathbf{x}$ as features, the model for $y_2$ uses $\mathbf{x}$ and $y_1$ as features and so on.) Passing label information between binary classifiers allows CC to take label dependencies into account. In the prediction stage, CC successively predicts the labels one at a time. The prediction results of the previous labels are used for predicting the next label in the chain.

This idea is extended to Probabilistic Classifier Chains (PCC) (Dembczyński et al., 2010). PCC explains CC using a probabilistic model. Specifically, the conditional joint distribution can be described as

$$P(y_1, ..., y_L\,|\,\mathbf{x}) = P(y_1\,|\,\mathbf{x}) \prod_{j=2}^{L} P(y_j\,|\,y_1, \ldots, y_{j-1}, \mathbf{x}) \tag{3.1}$$

and PCC estimates the probabilities $P(y_1\,|\,\mathbf{x}),\, P(y_2\,|\,\mathbf{x}, y_1), \ldots, P(y_L\,|\,\mathbf{x}, y_1, y_2, \ldots, y_{L-1})$.

PCC finds the label set that maximizes the right hand side of equation (3.1). However, there is no closed-form solution for finding the label set. A few different solutions have been suggested. Dembczyński et al. (2010) used an exhaustive search (ES) that considers all possible combinations. However, an exhaustive search may not be practical when $L$ is large, because the number of possible combinations $(2^L)$ increases exponentially. To overcome this problem, optimization strategies based on the uniform cost search (UCS) (Dembczyński, Waegeman and Hüllermeier, 2012) and the $A^*$ algorithm (Mena, Montañés, Quevedo and Del Coz, 2015) have been proposed. First, the estimated joint conditional probability may be represented by a probability binary tree. Then a search algorithm finds the optimal path (in our case, the path that gives the highest joint probability) from the root and the terminal node. Compared with ES, UCS substantially reduces the computational cost for PCC to reach the label set with the highest joint probability (Dembczyński et al., 2012).

In theory, when applying the product rule, the order of the categories $y_1, \ldots, y_L$ does not matter. For example, both $P(y_1\,|\,\mathbf{x})\,P(y_2\,|\,y_1, \mathbf{x})$ and $P(y_2\,|\,\mathbf{x})\,P(y_1\,|\,y_2, x)$ equal to $P(y_1, y_2\,|\,\mathbf{x})$. In practice, the two chains may lead to different estimates. This means the performance of PCC may be affected by the order of the labels in the chain.

To alleviate the influence of the category order, an ensembling approach (EPCC) (Dembczyński et al., 2010) that combines multiple probabilistic chains has been proposed. First $m$ PCC models are trained where each PCC model is based on a randomized order of the labels. In the prediction stage, the average conditional joint probability over the $m$ PCC models is computed for each possible label set. Then the predicted label set is the label set with the highest average predicted probability. Let $\hat{P}_j\,(\mathbf{Y}\,|\,\mathbf{x})$ be the conditional joint probability estimated by the $j^{\text{th}}$ PCC model. The ensemble strategy predicts the label set $\hat{\mathbf{Y}}$ such that

$$\hat{\mathbf{Y}} \;=\; \operatorname{argmax}_{\mathbf{Y}} \frac{\sum_{j=1}^{m} \hat{P}_j\,(\mathbf{Y}\,|\,\mathbf{x})}{m}.$$

Note that EPCC does not combine the predicted label sets but conditional joint probabilities. To find the highest average probability from $m$ PCC models, all individual probabilities are required and this forces us to use ES to compute the conditional joint probability for all $2^L$ label combinations from all $m$ PCC models. Hence, although EPCC reduces the problem of influence of label order, the method will not be useful if the problem deals with a large number of labels or when $m$ is large. To reduce the computational cost for combining multiple PCC models, we propose a new approach to ensembling the PCC models in the next section.

# 4 The majority-voted-based ensemble of PCC for semi-automated classification

The proposed method aims to ensemble multiple PCC models at much less computational cost. As mentioned in Section 3.2, the best label set (with the highest joint probability) for a single PCC can be found by a fast search strategy. In this paper, we use UCS, since the implementation is simple and the algorithm always finds the optimal solution. Using UCS, the proposed method obtains $\hat{\mathbf{Y}}_j\ (j = 1, \ldots, m)$, the label set predicted by the $j^{\text{th}}$ PCC model and $\hat{P}_j$, the estimated probability that $\hat{\mathbf{Y}}_j$ is the true label set. Among the $m$ predicted label sets, the proposed method chooses the most frequent label set for the final prediction. That is, $\hat{\mathbf{Y}} = \operatorname{mode}\left(\left\{\hat{\mathbf{Y}}_1, \ldots, \hat{\mathbf{Y}}_m\right\}\right)$. In case there are ties in the mode, we choose the label set whose averaged probability estimate is the highest.

Semi-automatic classification requires a score that measures how easy/hard the prediction is. Whether a text answer is classified automatically or manually is determined based on this score. Next, a score is proposed: Let $J$ be the set that contains all indices $j\ (1 \leq j \leq m)$ for which $\hat{\mathbf{Y}}_j$ is the most frequent one $\left(\text{i.e., } J = \left\{j : \hat{\mathbf{Y}}_j = \hat{\mathbf{Y}}\right\}\right)$. The proposed score for the prediction is

$$\theta = \left(\frac{\sum_{i \in J} \hat{P}_j}{|\,J\,|}\right)\left(\frac{|\,J\,|}{m}\right) \tag{4.1}$$

$$= \frac{\sum_{i \in J} \hat{P}_j}{m}. \tag{4.2}$$

The first factor of equation (4.1) is the average joint probability of the predicted label set. The second factor of equation (4.1) is the fraction of the PCC models that predict the predicted label set among the $m$ models. Multiplying the two components makes sense: a prediction may be more accurate if the (average) probability related to the chosen label set is high (the first factor) *and* more individual chain models vote for the same label set (the second component). We call this approach Majority-vote-based Ensemble of Probabilistic Classifier Chains (MEPCC). We later show empirically that combining the two factors indeed improves performance over just using a single factor. Table 4.1 illustrates an example for 5 labels $(L = 5)$ and 7 PCC models $(m = 7)$. The MEPCC approach stores the probability of one label set from each PCC model. Because MEPCC combines over the probabilities corresponding to the best label set from different PCC models, MEPCC can take advantage of the UCS (or any other) strategy. Note that a search strategy like UCS cannot be used for EPCC where all individual probabilities for all label combinations are required. More succinctly, MEPCC combines over the maximal probabilities of each PCC, whereas EPCC maximizes over the average probabilities, requiring evaluation of all individual probabilities. We summarize the procedure of MEPCC in Algorithm 1.

**Table 4.1**
**An example of the MEPCC classification of a single observation with $L = 5$ and $m = 7$**

| PCC model | Prediction | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $P(y_1, \ldots, y_5 \mid \mathbf{x})$ |
|---|---|---|---|---|---|---|---|
| 1 | $\hat{\mathbf{Y}}_1$ | 1 | 1 | 0 | 0 | 1 | 0.875 |
| 2 | $\hat{\mathbf{Y}}_2$ | 1 | 1 | 0 | 0 | 1 | 0.921 |
| 3 | $\hat{\mathbf{Y}}_3$ | 0 | 0 | 1 | 1 | 0 | 0.743 |
| 4 | $\hat{\mathbf{Y}}_4$ | 0 | 0 | 0 | 1 | 0 | 0.882 |
| 5 | $\hat{\mathbf{Y}}_5$ | 0 | 0 | 0 | 1 | 0 | 0.643 |
| 6 | $\hat{\mathbf{Y}}_6$ | 0 | 1 | 0 | 1 | 0 | 0.739 |
| 7 | $\hat{\mathbf{Y}}_7$ | 1 | 1 | 0 | 0 | 1 | 0.824 |
| final prediction | $\hat{\mathbf{Y}}$ | 1 | 1 | 0 | 0 | 1 | $\theta = \frac{0.875 + 0.921 + 0.824}{7} = 0.374$ |

**Algorithm 1.** The MEPCC algorithm

```
Input: Number of models $m$, an instance vector $\mathbf{x}$, corresponding PCC models $h_j$, the uniform cost search algorithm $U$
for $j = 1$ to $m$ do
        (a) Using $h_j$ and $U$, obtain $\hat{\mathbf{Y}}_j = \text{argmax}_{\mathbf{Y}} P(\mathbf{Y} \mid \mathbf{x})$
        (b) Store $\hat{P}_j = P(\hat{\mathbf{Y}}_j \mid \mathbf{x})$
end for
Obtain the label set $\hat{\mathbf{Y}} = \text{mode}(\{\hat{\mathbf{Y}}_1, \ldots, \hat{\mathbf{Y}}_m\})$
Obtain $J = \{j : \hat{\mathbf{Y}}_j = \hat{\mathbf{Y}}\}$
Obtain the score $\theta = \frac{\sum_{i \in J} \hat{P}_j}{m}$
Return $\hat{\mathbf{Y}}$ and $\theta$
```

# 5 Experiments

## 5.1 Data

We evaluated the performance of the MEPCC algorithm on three different data sets: Civil disobedience, Immigrant and Happy data (the Happy data are available upon request by contacting Marika Wenemark marika.wenemark@liu.se. The Immigrant and Civil Disobedience data are available from the GESIS Datorium http://dx.doi.org/10.7802/1795). For each data set, an open-ended question was asked to the respondents and their answers have been coded manually with possibly multiple labels.

The Civil data set was collected to study cross-cultural equivalence about Civil disobedience. Behr, Braun, Kaczmirek and Bandilla (2014) first asked respondents a closed-ended question from the ISSP (ISSP Research Group, 2012) How important is it that citizens may engage in acts of Civil disobedience when they oppose government actions? (Not at all important 1 − Very important 7). The respondents were then asked: What ideas do you associate with the phrase “Civil disobedience”? Please give examples. Answers were classified into 12 labels: non-productive, violence, disturbances, peaceful, listing activities, breadth of actions, breaking law, breaking rules, government:dissatisfaction, government:deep rift, copy/paste from the Internet, other. The survey data were collected in different languages and we use a merged data set (Spanish, German and Danish) that contains 1,029 observations.

The Immigrant data set was collected to study cross-national equivalence of measures of xenophobia. In the 2003 International Social Survey Program (ISSP) on National Identity, the questionnaire contained four statements regarding beliefs on Immigrants such as Immigrants take jobs from people who were born in Germany. After rating each statement, respondents were asked to answer to an open-ended question: Which type of Immigrants were you thinking of when you answered the question? The previous statement was: [text of the corresponding item]. Braun, Behr and Kaczmirek (2013) classified answers into 14 labels: non-productive, positive, negative, neutral/work, general, Muslim countries, eastern European, Asia, ex-Yugoslavia, EU15, sub Sahara, Sinti/Roma, legal/illegal, other. In this article, we use 1,006 observations from the German survey.

The Happy data set was collected to study the relationship between positive factors and mental health and care needs. Wenemark, Borgstedt-Risberg, Garvin, Dahlin, Jusufbegovic, Gamme, Johansson and Bjrn (2018) asked respondents “Name some positive things in your life, that are uplifting or make you Happy: (you may write several things)”. Answers were classified into 13 labels: nothing, relationships (family or romantic), working/studying, health, self-esteem, joy/happiness, well-being: drinking/eating/drugs/sex, spirituality, money, nature, hobbies, culture, and exercise. The data set contains 2,350 observations.

Table 5.1 contains summary statistics about the three data sets.

**Table 5.1**
**Summary statistics of data sets: number of total observations, features and labels and average number of relevant labels, and percentage of observations that are associated with more than one label $(\mathbf{P}_{|L|>1})$**

| Data | # observations | # features | L | av. # of labels | $\mathbf{P}_{\vert L\vert >1}$ |
|---|---|---|---|---|---|
| Civil | 1,029 | 305 | 12 | 1.15 | 13.80% |
| Immigrant | 1,006 | 273 | 14 | 1.19 | 13.72% |
| Happy | 2,350 | 492 | 13 | 2.77 | 87.40% |

## 5.2 Experimental setup

We compared the proposed MEPCC method against BR and LP and PCC. For PCC, we used the uniform search to reach a predicted label set and the estimated probability of equation (3.1) for the confidence score of the prediction. EPCC was not included in the comparison because its computational cost makes it infeasible for prediction for our data sets. (In our experiment on the Immigrant data with 14 labels, running the exhaustive search for PCC $(m=1)$ for a single prediction took a single computer (Intel Core i7 CPU with 8GB RAM) over 30 minutes. This implies that predicting 200 observations using EPCC $(m=10)$ would take more than 1,000 hours.) Support vector machines (SVM) (Vapnik, 2000) were used as the base classifier on unscaled variables with a linear kernel and tuning parameter $C=1$. For probabilistic output, the SVM scores were converted into probabilities using Platt's method (Platt, 2000). The analysis was conducted in $R$ (R Core Team, 2014) using the $e1071$ package (Meyer, Dimitriadou, Hornik, Weingessel and Leisch, 2014) for SVM.

For each data set, 5-fold cross validation (CV) was performed. That is, we randomly divided the data into five equal-sized parts and used the first four parts as the training data and the last part as the test data. Performance evaluation is only made on the test data. Each of the five parts were used as test data and the results were averaged.

## 5.3 Performance of the MEPCC approach

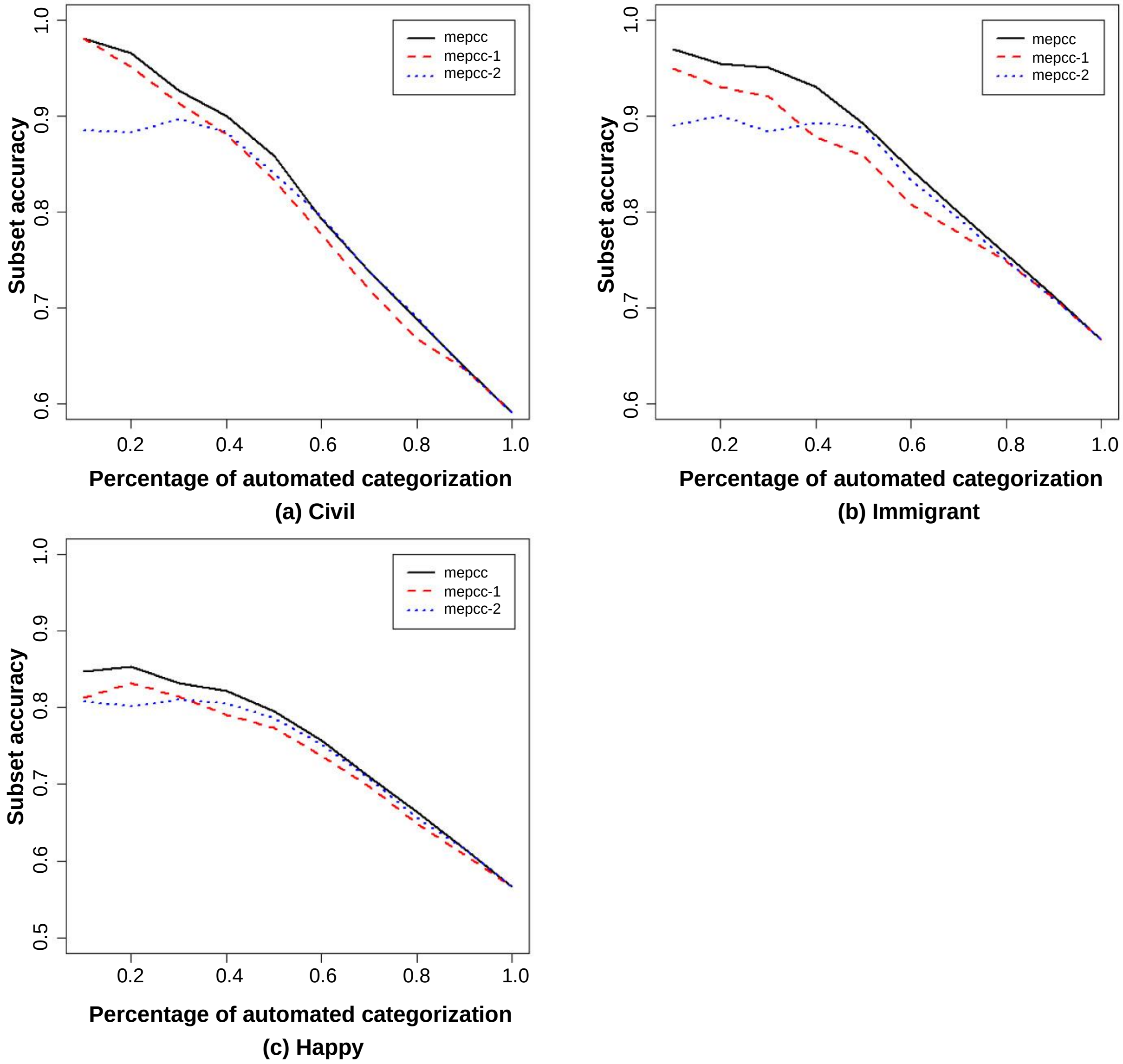
We first investigated the performance of the MEPCC. The score in equation (4.1) has two components. To demonstrate that both components are helpful, we evaluate the proposed score as well as two different scores where one of the components is missing. That is, we compared the MEPCC with three different scores $\theta$, $\theta_1$ and $\theta_2$:

$$\text{(MEPCC)}\ \theta = \left(\frac{\sum_{i\in J}\mathbf{P}_j}{|J|}\right)\left(\frac{|J|}{m}\right)$$

$$\text{(MEPCC} - 1) \quad \theta_1 = \left(\frac{\sum_{i\in J}\mathbf{P}_j}{|J|}\right)$$

$$\text{(MEPCC} - 2) \quad \theta_2 = \left(\frac{|J|}{m}\right).$$

Prioritizing the text answers based on $\theta_2$ results in many ties. The tied answers were randomly reordered to be able to calculate subset accuracy at each production rate. Figure 5.1 shows the subset accuracy of each approach as a function of the production rate. The text answers with higher scores were classified first. For example, production rate 0.2 means only 20% of the test data with the highest scores were classified automatically by the models. When the production rate equals 1, there was no difference between the MEPCC models because the predicted label sets are always the same. The difference is how they prioritize the text answers from the easiest-to-classify to the hardest-to-classify answers. When the production rate was less than 1, MEPCC outperformed MEPCC-1 and MEPCC-2 for all three data. The results show that both components in equation (4.1) were helpful for prioritizing the observations.



**(a) Civil**

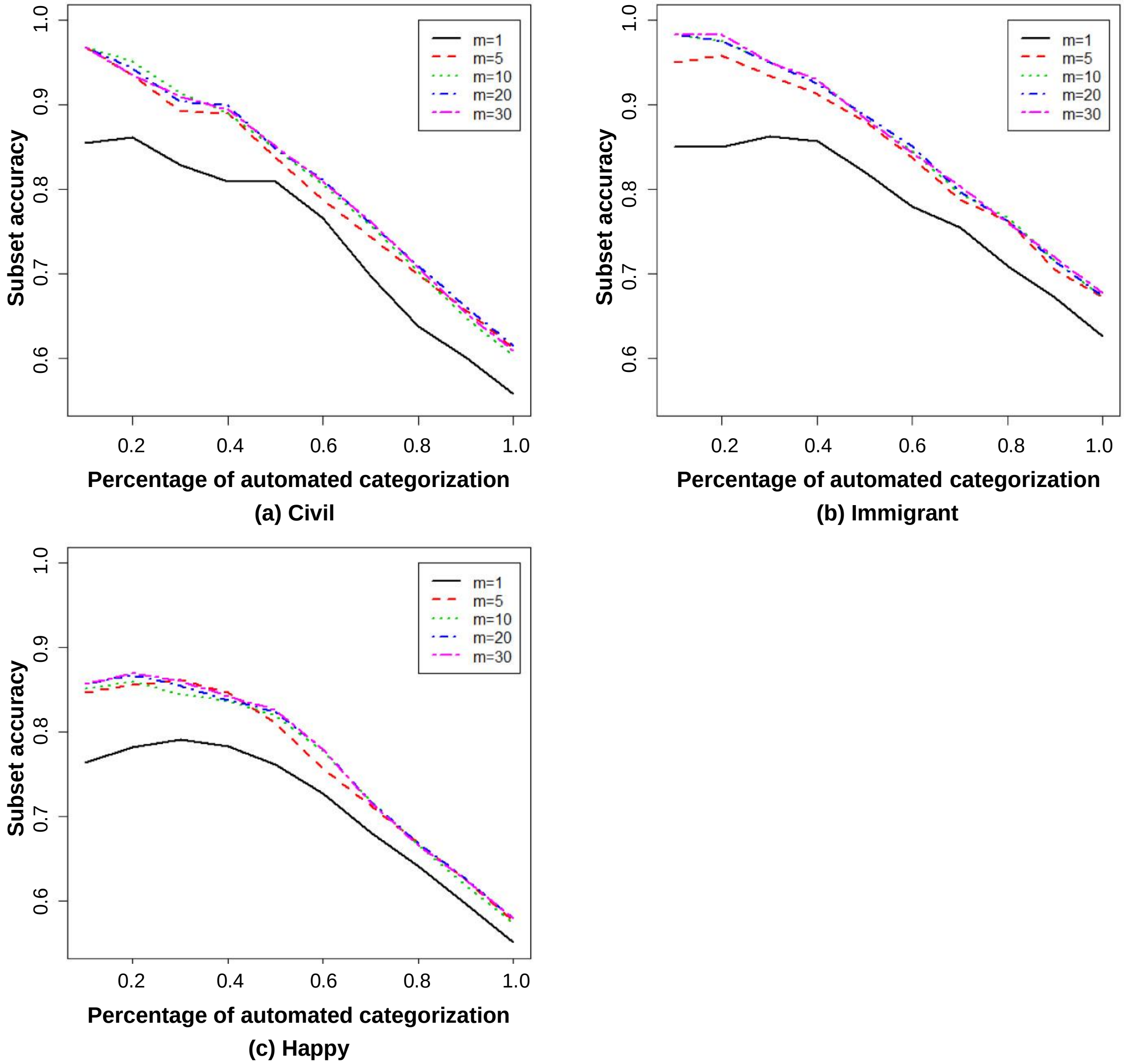
**(b) Immigrant**

**(c) Happy**

**Figure 5.1 Subset accuracy of three variations on MEPCC as a function of production rate.**

## 5.4 Effect of the number of PCC models

We then investigated to what extent the number of PCC models affects the predictive performance of MEPCC. Figure 5.2 shows the performance of MEPCC for different number of PCC models $(m)$. When $m$ was low, increasing $m$ led to huge improvement of the subset accuracy of MEPCC. However, once there were enough PCC models (e.g., $m = 10$), adding more PCC models did not improve the subset accuracy. The empirical results show that MEPCC does not require many PCC models for performing well.



(a) Civil

(b) Immigrant

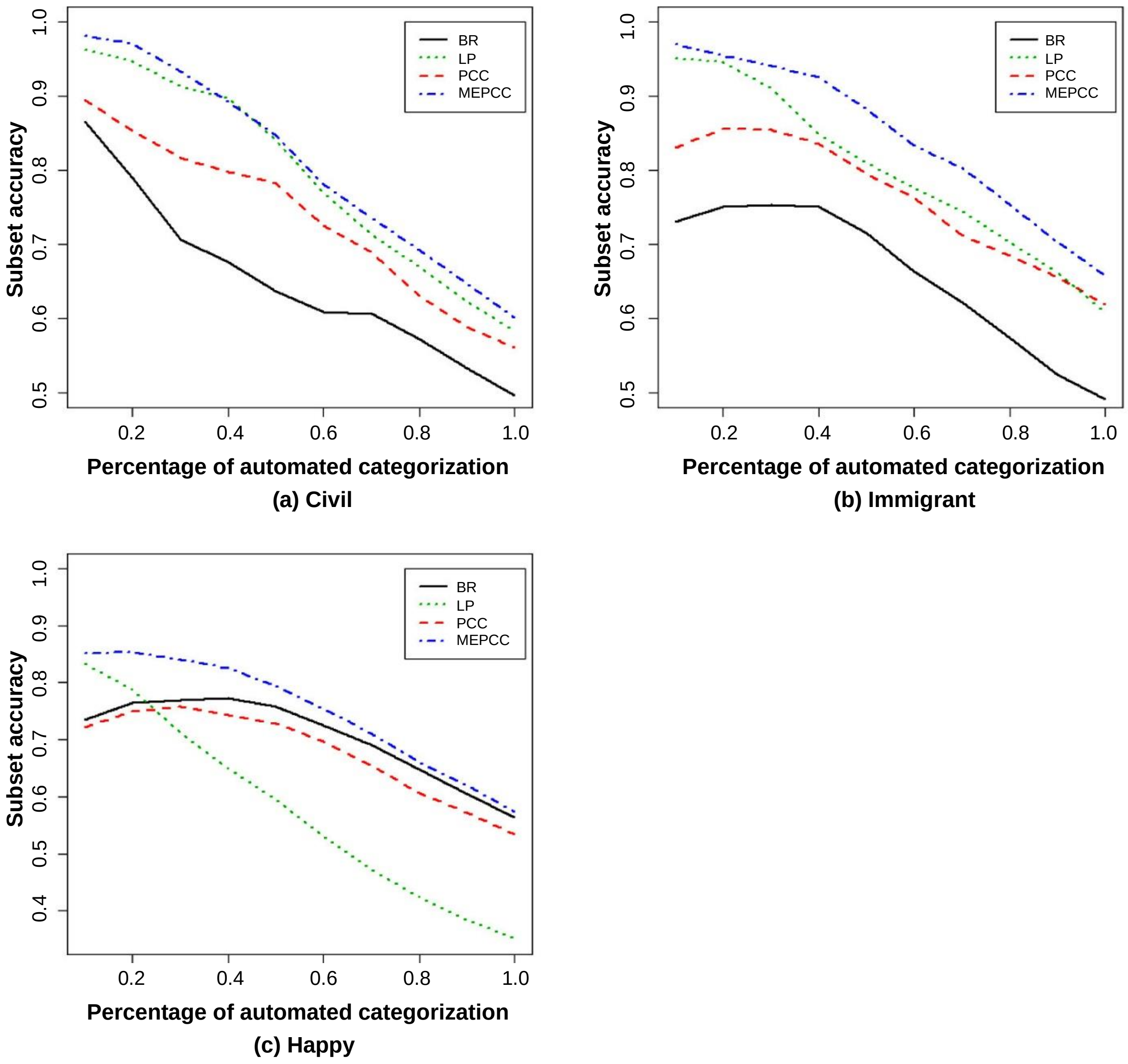
(c) Happy

**Figure 5.2 The effect of the number of PCC models (m) used for MEPCC.**

## 5.5 Comparison with other methods

At last we investigated the performance of MEPCC $(m = 10)$ compared to the established methods (BR, LP and PCC). For all methods, a production rate of x% refers to the x% of the data that have the highest score. MEPCC used $\theta$ as a score, while each of the other approaches used the probability of the predicted label set estimated by that method. Note when $m = 1$, MEPCC and PCC are identical; the score $\theta$ coincides with the probability of the label set predicted by PCC.

Figures 5.3 and 5.4 illustrate the respective subset accuracy and Hamming loss for the different methods as a function of the production rate on the Happy, Immigrant and Civil data. For the Immigrant and Happy data, the highest subset accuracy at most production rates was obtained by MEPCC. For the Civil data, MEPCC and LP performed the best. In terms of Hamming loss, MEPCC achieved the lowest error at most production rates for all data.



(a) Civil

(b) Immigrant

(c) Happy

**Figure 5.3 Semi-automated result (subset accuracy) for the three data from the 5-fold cross validation.**

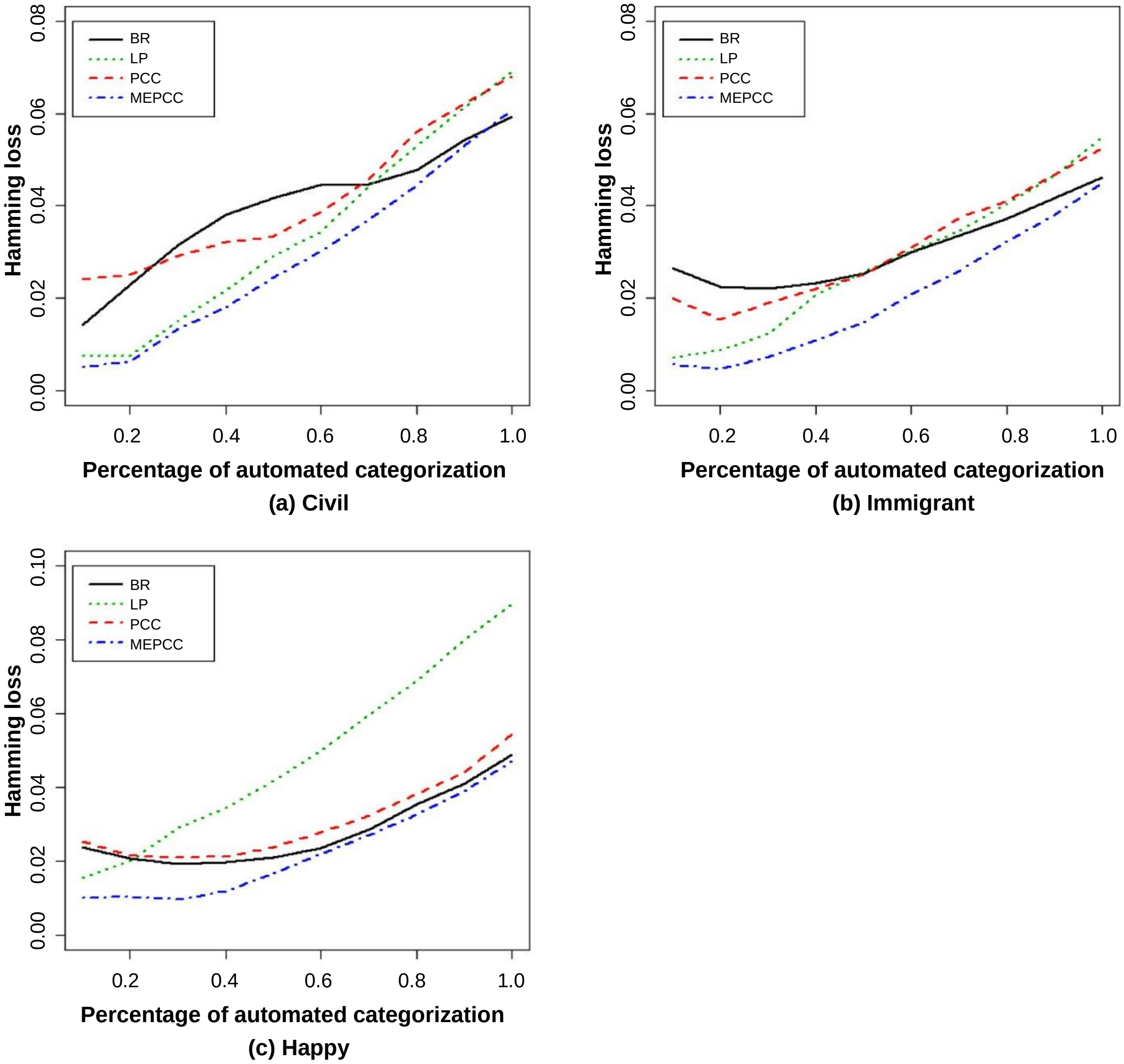


**Figure 5.4 Semi-automated result (Hamming loss) for the three data from the 5-fold cross validation.**

Next, we consider the performance of each method given target predicted accuracy values. To decide the fraction of automatic categorization, a practitioner will typically set a threshold probability above which texts are coded automatically. For MEPCC, the relationship between true accuracy and the confidence score $(\theta)$ were estimated via cross-validation on the training data. We used Platt's scaling to convert the confidence scores into probability outputs. Since Platt's scaling could improve the level of calibration (Niculescu-Mizil and Caruana, 2005), the same technique was also applied to BR, LP and PCC.

Table 5.2 illustrates the tradeoff between the percentages of automated prediction and the corresponding subset accuracy of each method as a function of different thresholds. The threshold refers to the minimum predicted subset accuracy required for automated prediction. The minimum predicted subset accuracy helps us decide which text answers should be classified automatically and which should be classified manually. For example, if the client decides that at least 80% accuracy is required for automated classification, then approximately 39.3% of the Civil data, 42.5% of the Immigrant data, and 27.6% of the Happy data can be classified automatically by MEPCC with subset accuracy 0.891, 0.916 and 0.857, respectively. Note that this is a huge improvement compared to applying BR that could only automatically classify 9.3% of the Civil data, 12.8% of the Immigrant data, and 8.7% of the Happy data with lower subset accuracies. Table 5.3 shows the relationship between predicted and actual accuracy by aggregating to ranges of predictions for each method and data set. For MEPCC the actual accuracy is within the range of the predicted accuracy in most cases, much better than for the other methods.

**Table 5.2**
**Semi-automated result for the three data at different decision thresholds. P represents the percentage of automated predictions and SA represents the subset accuracy for the automated prediction results**

| Data | Threshold | BR | | LP | | PCC | | MEPCC | |
|---|---|---|---|---|---|---|---|---|---|
| | | P | SA | P | SA | P | SA | P | SA |
| Civil | 0.9 | 0.7% | 0.667 | 16.5% | 0.967 | 0.0% | NA | 13.0% | 0.978 |
| | 0.8 | 9.3% | 0.893 | 34.3% | 0.898 | 15.1% | 0.787 | 39.3% | 0.891 |
| | 0.7 | 18.4% | 0.846 | 46.6% | 0.852 | 36.4% | 0.817 | 45.8% | 0.860 |
| | 0.6 | 25.4% | 0.768 | 50.6% | 0.831 | 52.1% | 0.771 | 52.9% | 0.820 |
| Immigrant | 0.9 | 3.7% | 0.858 | 11.1% | 0.959 | 1.3% | 0.558 | 31.5% | 0.947 |
| | 0.8 | 12.8% | 0.779 | 30.4% | 0.890 | 27.7% | 0.859 | 42.5% | 0.916 |
| | 0.7 | 26.6% | 0.743 | 38.6% | 0.863 | 42.4% | 0.829 | 55.1% | 0.862 |
| | 0.6 | 41.7% | 0.715 | 53.6% | 0.806 | 50.5% | 0.795 | 62.7% | 0.839 |
| Happy | 0.9 | 1.3% | 0.592 | 8.9% | 0.850 | 0.1% | 0.750 | 1.0% | 0.830 |
| | 0.8 | 8.7% | 0.734 | 14.3% | 0.802 | 7.2% | 0.726 | 27.6% | 0.857 |
| | 0.7 | 32.8% | 0.776 | 17.7% | 0.793 | 29.9% | 0.767 | 43.7% | 0.817 |
| | 0.6 | 53.2% | 0.745 | 22.2% | 0.761 | 49.2% | 0.744 | 52.0% | 0.790 |

**Table 5.3**
**Semi-automated result for the three data at different ranges of thresholds. P represents the percentage of automated predictions and SA represents the subset accuracy for the automated prediction results**

| Data | Predicted accuracy | BR | | LP | | PCC | | MEPCC | |
|---|---|---|---|---|---|---|---|---|---|
| | | P | SA | P | SA | P | SA | P | SA |
| Civil | [0.9, 1.0] | 0.7% | 0.667 | 16.5% | 0.967 | 0.0% | NA | 13.0% | 0.978 |
| | [0.8, 0.9) | 8.7% | 0.896 | 17.8% | 0.834 | 15.1% | 0.787 | 26.2% | 0.846 |
| | [0.7, 0.8) | 9.0% | 0.769 | 12.2% | 0.710 | 21.3% | 0.828 | 6.5% | 0.681 |
| | [0.6, 0.7) | 7.0% | 0.566 | 4.1% | 0.584 | 15.7% | 0.655 | 7.1% | 0.563 |
| Immigrant | [0.9, 1.0] | 3.7% | 0.858 | 11.1% | 0.959 | 1.3% | 0.558 | 31.5% | 0.947 |
| | [0.8, 0.9) | 9.1% | 0.750 | 19.3% | 0.843 | 26.4% | 0.869 | 11.0% | 0.829 |
| | [0.7, 0.8) | 13.8% | 0.710 | 8.2% | 0.747 | 14.7% | 0.757 | 12.5% | 0.688 |
| | [0.6, 0.7) | 15.1% | 0.602 | 15.0% | 0.659 | 8.1% | 0.623 | 7.7% | 0.670 |
| Happy | [0.9, 1.0] | 1.3% | 0.592 | 8.9% | 0.850 | 0.1% | 0.750 | 1.0% | 0.830 |
| | [0.8, 0.9) | 7.4% | 0.755 | 5.4% | 0.717 | 7.1% | 0.730 | 26.5% | 0.858 |
| | [0.7, 0.8) | 24.0% | 0.792 | 3.4% | 0.751 | 22.7% | 0.779 | 16.2% | 0.749 |
| | [0.6, 0.7) | 20.4% | 0.693 | 4.6% | 0.615 | 19.3% | 0.703 | 8.3% | 0.647 |

Table 5.4 shows the runtime of each method for training the model and predicting all instances in test data (Intel Core i7 CPU with 8GB RAM). Unsurprisingly, the runtime of MEPCC at $m = 10$ is roughly 10 times of that of PCC in both of the training and prediction stages.

**Table 5.4**
**Runtime (in seconds) of each method for the three data**

| Data | Stage | BR | LP | PCC | MEPCC |
|---|---|---|---|---|---|
| Civil | Train | 1.688 | 0.641 | 1.128 | 11.787 |
| | Prediction | 0.269 | 0.044 | 37.142 | 374.611 |
| Immigrant | Train | 1.363 | 0.510 | 0.894 | 8.724 |
| | Prediction | 0.200 | 0.056 | 35.369 | 334.075 |
| Happy | Train | 11.160 | 16.164 | 7.371 | 78.293 |
| | Prediction | 0.567 | 3.691 | 177.847 | 1,746.529 |

# 6 Discussion

Using three examples, we have investigated several approaches for automated classification for any desired production rate when data are multi-labeled. In terms of subset accuracy and Hamming loss, the proposed method, MEPCC, achieved the best performance at most production rates in all three data sets.

There were trade-offs between the prediction performance and the production rate for all methods. At low production rates, high subset accuracy and low Hamming loss were achieved for a small number of easy-to-classify answers. However, accuracy (loss) tended to decrease (increase) as more difficult answers were included (i.e., production rate increased).

Either subset accuracy or production rate can be set at a target rate which determines the second measure. For example, targeting 80% minimum subset accuracy for an automated prediction, MEPCC categorizes 39.3% of the Civil data, 42.5% of the Immigrant data, and 27.6% of the Happy data automatically. Such a reduction is considerable. In an applied research environment, reducing the need for manual coding in a data set with 5,000 observations, a reduction by 50% may save several weeks of coding time. If production rate is fix at 80%, MEPCC could achieve a subset accuracy of 70% (Civil), 75% (Immigrant), and 68% (Happy).

The Hamming loss represents the fraction of misclassified labels. Figure 5.4 shows that the improvement of MEPCC over BR was quite noticeable at lower production rates but relatively small at 100% production rate.

MEPCC outperformed PCC at most production rates on all three data. This shows that combining multiple PCC models substantially improves the performance. As can be seen from Figure 5.2, even combining 5 models resulted in a substantial improvement throughout the whole range of production rate. The difference tended to be greater at lower production rates. This means MEPCC is even more preferred for semi-automated classification, where a high accuracy is required rather than a high production rate. The performance of MEPCC converged as $m$ increased in all three data sets. The difference between the

MEPCC models were negligibly small when should be an equation was larger than 10. This is a desirable result in practice because employing too many PCC models for an ensemble model is unnecessary.

For all three data we found that the proposed method was not sensitive to the choice of the search algorithm for each PCC model (results and figures not shown). That is, the classification results of MEPCC with the uniform cost search were similar to those with the greedy search. While the proposed method uses the uniform cost search, the greedy approach may also be considered especially when the fast prediction time matters.

Figure 5.3 shows LP beats BR for the Civil and Immigrant data sets and BR beats LP for the Happy data set with respect to subset accuracy. We see two reasons: 1) LP performed well when the number of unique label sets was relatively small (Civil: 39, Immigrant: 59). However, the performance of LP was not effective but less well for the Happy data where the number of unique label sets was large (346). 2) BR does not take into account correlations among the labels. BR beat LP where bivariate label correlation were low (Happy data) and LP beat BR where bivariate label correlations were larger (Civil and Immigrant Data). Compared to BR and LP, MEPCC seems to be robust to those aspects (the number of unique label sets and the magnitude of label correlations).

The semi-automatic procedure introduced here works best in repeated survey questions where results from previous waves have been labeled or for one-off questions where the sample size is large. How large should the training data be? We have used 5-fold cross-validation to evaluate the algorithm, but cross-validation is not appropriate in a production environment. If the question was asked in a previous wave, train the algorithm on all labeled data from all previous waves. If not, set a “sufficiently large” number of texts aside for labeling and training, and use the semi-automatic procedure on the remainder of the data. How large “sufficiently large” is depends on the task at hand. For single labeling tasks we have found that often 500 training samples are sufficient (Schonlau and Couper, 2016). There is a tradeoff: a larger data set predicts more accurately but also reduces the scope for time savings as fewer unlabeled observations remain. Under reasonable assumptions, Schonlau and Couper (2016) suggested human coding time savings for a single-label semi-automatic coding procedure attempting to code 1,000 (9,500) texts might be 14 (133) hours. 133 hours is equivalent to 16.6 eight-hour working days. Whether those time savings are large enough to warrant implementation of a semi-automatic procedure may be best decided with knowledge of the specific task and in the context of the specific production environment.

If some label combinations cannot occur in individual data sets, such constraints on label combinations may be added. For example, for the Happy data, if the label “nothing” is turned on all other labels must be turned off. Knowing that “nothing” is incompatible with other labels requires some domain expertise. It would be straightforward to modify the algorithm to accommodate this constraint. Of course, all methods except BR already exploit dependencies between labels; implementing this constraint may not affect performance very much. We did not implement such constraints in this article to avoid the appearance of the algorithms heavily relying on the constraints.

Limitations of this work include that the experimental study was conducted using three text data sets only. While there is no guarantee that performance will be equally good on other data sets, data used in this paper consider different topics in different languages, which increases the appeal of MEPCC. Also, all of the multi-label algorithms in this article used the same base learner (SVM) for classification. While SVM is one of the best performing approaches, other learning methods that produce probability outcomes could be chosen.

In conclusion, we investigated semi-automated classification for open-ended questions when the data are multi-labelled using existing multi-label algorithms. We have proposed a new algorithm for semi-automatic classification that effectively combines multiple PCC models. The experimental results on three different example data show that the proposed approach outperforms BR, LP and PCC in terms of subset accuracy and Hamming loss at most production rates. Although we focused on survey data from open-ended questions, the proposed approach can also be applied to other types of multi-label data when semi-automated classification is desired. A comprehensive analysis encompassing a variety of data in the context of semi-automated classification deserves further investigation.

# References

Behr, D., Braun, M., Kaczmirek, L. and Bandilla, W. (2014). Item comparability in crossnational surveys: Results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity*, 48(1), 127-148.

Braun, M., Behr, D. and Kaczmirek, L. (2013). Assessing cross-national equivalence of measures of xenophobia: Evidence from probing in web surveys. *International Journal of Public Opinion Research*, 25(3), 383-395.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Dembczyński, K., Cheng, W. and Hüllermeier, E. (2010). Bayes optimal multilabel classification via probabilistic classifer chains. *Proceedings of the 27th International Conference on Machine Learning*, 279-286.

Dembczyński, K., Waegeman, W. and Hüllermeier, E. (2012). An analysis of chaining in multi-label classification. In *Frontiers in Artificial Intelligence and Applications*, (Eds., L. De Raedt, C. Bessiere, D. Dubois, P. Doherty, P. Frasconi, F. Heintz and P. Lucas), 242, 294-299. IOS Press.

Guenther, N., and Schonlau, M. (2016). Support vector machines. *The Stata Journal*, 16(4), 917-937.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. and Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, 33(1), 101-122.

ISSP Research Group (2012). International social survey programme: Citizenship - ISSP 2004. GESIS data archive, Cologne. ZA3950 data file version 1.3.0, https://doi.org/10.4232/1.11372.

Manning, C., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*, Chapter 2.2. Cambridge, England: Cambridge University Press.

Matthews, P., Kyriakopoulos, G. and Holcekova, M. (2018). Machine learning and verbatim survey responses: Classification of criminal offences in the crime survey for England and Wales. Paper presented at BigSurv18, Barcelona, Spain.

Mena, D., Montañés, E., Quevedo, J.R. and Del Coz, J.J. (2015). Using A* for inference in probabilistic classifer chains. *Proceedings of the 24th International Conference on Artificial Intelligence*, 3707-3713. AAAI Press.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. (2014). *e1071: Misc Functions of The Department of Statistics, TU Wien*. http://CRAN.R-project.org/package=e1071.

Niculescu-Mizil, A., and Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, New York, NY, U.S.A., 625-632. ACM.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, (Eds., A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans), 61-74. MIT Press.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Read, J., Pfahringer, B., Holmes, G. and Frank, E. (2009). Classifier chains for multi-label classification. In *Machine Learning and Knowledge Discovery in Databases*, (Eds., W. Buntine, M. Grobelnik, D. Mladenić and J. Shawe-Taylor), 254-269. Springer.

Read, J., Pfahringer, B., Holmes, G. and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3), 333-359.

Schonlau, M., and Couper, M.P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, 10(2), 143-152.

Schonlau, M., Guenther, N. and Sucholutsky, I. (2017). Text mining using ngram variables. *The Stata Journal*, 17(4), 866-881.

Schonlau, M., Gweon, H. and Wenemark, M. (2019). Automatic classification of open-ended questions: Check-all-that-apply questions. *Social Science Computer Review*. First published online August 20, 2019 (to appear in a future issue). https://journals.sagepub.com/doi/full/10.1177/0894439319869210.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47.

Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory*. 2nd Edition. Springer.

Wenemark, M., Borgstedt-Risberg, M., Garvin, P., Dahlin, S., Jusufbegovic, J., Gamme, C., Johansson, V. and Bjrn, E. (2018). Psykisk hlsa i sydstra sjukvrdsregionen: En kartlggning av sjlvskattad psykisk hlsa i jnkping. Kalmar och stergtlands ln hsten 2015/16. Retrieved from https://vardgivarwebb.regionostergotland.se/pages/285382/Psykisk_halsa_syostra_sjukvarsregionen.pdf.

Ye, C., Medway, R. and Kelley, C. (2018). Natural language processing for open-ended survey questions. Paper presented at BigSurv18, Barcelona, Spain.

# ACKNOWLEDGEMENTS

*Survey Methodology* wishes to thank the following people who have provided help or served as referees for one or more papers during 2020.

- P.G. Andersson, *Stockholm University*
- J.-F. Beaumont, *Statistics Canada*
- E. Ben-David, *US Census Bureau*
- M. Brick, *Westat Inc.*
- J. Brown, *University of Technology Sydney*
- S. Cai, *Carleton University*
- P.J. Cantwell, *U.S. Census Bureau*
- G. Chatrchi, *Statistics Canada*
- R. Clark, *Australian National University*
- S. Chatterjee, *University of Minnesota*
- G. Chauvet, *École nationale de la statistique et de l'analyse de l'information*
- S. Chen, *University of Oklahoma Health Sciences Center*
- J. Chipperfield, *Australian Bureau of Statistics*
- M. Couper, *University of Michigan*
- J. Dever, *RTI International*
- D. Dolson, *Statistics Canada*
- J.L. Eltinge, *U.S. Bureau of Labor Statistics*
- N. English, *National Opinion Research Center*
- H. Fernandes, *Federal University of Paraíba*
- W.A. Fuller, *Iowa State University*
- J. Gambino, *Statistics Canada*
- C. Girard, *Statistics Canada*
- W. Gross, *University of Wollongong*
- D. Haziza, *Université de Montréal*
- P. Han, *University of Michigan*
- M.A. Hidiroglou, *Statistics Canada*
- M. Hirose, *Kyushu University*
- B. Hulliger, *University of Applied and Arts Sciences Northwestern Switzerland*
- K. Irimata, *US Census Bureau*
- S. Jobe, *Bureau of Labor Statistics*
- D. Judkins, *ABT Associates Inc Bethesda*
- P. Kelly, *Statistics Canada*
- J. Kim, *Iowa State University*
- S. Kolenikov, *ABT Associates Inc*
- P. Kott, *RTI International*
- S. Krieg, *Statistics Netherlands*
- P. Lahiri, *University of Maryland*
- E. Lesage, *L'Institut national de la statistique et des études économiques*
- B. Liu, *National Cancer Institute*
- D. Marella, *Roma Tre University*
- J.M. Marín, *Universidad Carlos III de Madrid*
- K. McConville, *Reed College*
- I. Molina, *Universidad Carlos III de Madrid*
- D. Nedyalkova, *Swiss Federal Statistical Office*
- J. Opsomer, *Westat Inc*
- D. Pfeffermann, *University of Southampton*
- J. Preston, *Australian Bureau of Statistics*
- M. Provost, *Statistics Canada*
- J.N.K. Rao, *Carleton University*
- L.-P. Rivest, *Université Laval*
- E. Rocco, *University of Florence*
- F. Scheuren, *National Opinion Research Center*
- P.L.N.D. Silva, *Escola Nacional de Ciências Estatísticas*
- P. Smith, *University of Southampton*
- D. Steel, *University of Wollongong*
- T. Tarray, *Islamic University of Science and Technology*
- M. Torabi, *University of Manitoba*
- D. Toth, *U.S. Bureau of Labor Statistics*
- A.-A. Vallée, *Université Laval*
- J. van den Brakel, *Statistics Netherlands*
- A. Wakeel, *COMSATS Institute of Information Technology*
- C. Wu, *University of Waterloo*
- A. Zaslavsky, *Harvard University*
- L.-C. Zhang, *University of Southampton*

Acknowledgements are also due to those who assisted during the production of the 2020 issues: Céline Ethier and Cynthia Bocci of International Cooperation and Methodology Innovation Centre; Joana Bérubé of Economic Statistics Methods Division; Gilbert Côté of Official Release, Language Services and Respondent Communications Division; the team from Dissemination Division, in particular: Chantal Chalifoux, Kathy Charbonneau, Christina Jaworski, Joseph Prince, Travis Robinson, Beth Chepita, Jenna Waite and Michèle Lanoue as well as our partners in the Communications Division

# ANNOUNCEMENTS

## Nominations Sought for the 2022 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honor of the late Joseph Waksberg to recognize his outstanding contributions to survey statistics and methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey statistics and methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium and give the 2022 Waksberg Invited Address at the Statistics Canada Symposium, expected to be held in the autumn of 2022. The paper will be published in an upcoming issue of *Survey Methodology* (Targeted for December 2022).

The author of the 2022 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*. **Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2021 to the chair of the committee, Jean Opsomer, email: JeanOpsomer@westat.com**.

Previous Waksberg Award honorees and their invited papers are:

2021 Sharon **Lohr**, Manuscript topic under consideration.

2020 Roger **Tourangeau**, "Science and survey management", *Survey Methodology* (expected for vol. 47, 1).

2019 Chris **Skinner**.

2018 Jean-Claude **Deville**, "De la pratique à la théorie : l'exemple du calage à poids bornés". 10ème Colloque francophone sur les sondages, Université Lumière Lyon 2.

2017 Donald B. **Rubin**, "Conditional calibration and the sage statistician". *Survey Methodology*, vol. 45, 2, 187-198.

2016 Don A. **Dillman**, "The promise and challenge of pushing respondents to the Web in mixed-mode surveys". *Survey Methodology*, vol. 43, 1, 3-30.

2015 Robert M. **Groves**, "Towards a quality framework for blends of designed and organic data". Proceedings: *Symposium 2016, Growth in Statistical Information: Challenges and Benefits*.

2014 Constance F. **Citro**, "From multiple modes for surveys to multiple data sources for estimates". *Survey Methodology*, vol. 40, 2, 137-161.

2013 Ken **Brewer**, "Three controversies in the history of survey sampling". *Survey Methodology*, vol. 39, 2, 249-262.

2012 Lars **Lyberg**, "Survey quality". *Survey Methodology*, vol. 38, 2, 107-130.

2011 Danny **Pfeffermann**, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". *Survey Methodology*, vol. 37, 2, 115-136.

2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.

2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.

2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.

2007 **Carl-Erik Särndal,** "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.

2006 Alastair **Scott**, “Population-based case control studies”. *Survey Methodology*, vol. 32, 2, 123-132.

2005 J.N.K. **Rao**, “Interplay between sample survey theory and practice: An appraisal”. *Survey Methodology*, vol. 31, 2, 117-138.

2004 Norman M. **Bradburn**, “Understanding the question-answer process”. *Survey Methodology*, vol. 30, 1, 5-15.

2003 David **Holt**, “Methodological issues in the development and use of statistical indicators for international comparisons”. *Survey Methodology*, vol. 29, 1, 5-17.

2002 Wayne A. **Fuller**, “Regression estimation for survey samples”. *Survey Methodology*, vol. 28, 1, 5-23.

2001 Gad **Nathan**, “Telesurvey methodologies for household surveys – A review and some thoughts for the future”. *Survey Methodology*, vol. 27, 1, 7-31.

## Members of the Waksberg Paper Selection Committee (2020-2021)

Jean Opsomer, *Westat* (Chair)
Jack Gambino, *Statistics Canada*
Giovanna Ranalli, *University of Perugia*
Elizabeth Stuart, *John Hopkins Bloomberg School of Public Health*

**Past Chairs**:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)
Elizabeth A. Martin (2010 - 2011)
Mary E. Thompson (2011 - 2012)
Steve Heeringa (2012 - 2013)
Cynthia Clark (2013 - 2014)
Louis-Paul Rivest (2014 - 2015)
Tommy Wright (2015 - 2016)
Kirk Wolter (2016 - 2017)
Danny Pfeffermann (2017 - 2018)
Mike Hidiroglou (2018 - 2019)
Robert E. Fay (2019 - 2020)

# JOURNAL OF OFFICIAL STATISTICS

**An International Review Published by Statistics Sweden**

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents
## Volume 36, No. 2, June 2020

All inquires about submissions and subscriptions should be directed to jos@scb.se

# JOURNAL OF OFFICIAL STATISTICS

**An International Review Published by Statistics Sweden**

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents
## Volume 36, No. 3, September 2020

All inquires about submissions and subscriptions should be directed to jos@scb.se

**The Canadian Journal of Statistics** **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

## Volume 48, No. 1, March/mars 2020

### Editorial

### Original Articles

### Acknowledgement

**The Canadian Journal of Statistics** **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

## Volume 48, No. 2, June/juin 2020

### Original Articles

# GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). Before submitting the article, please examine a recent issue of *Survey Methodology* (Vol. 39, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word with MathType for the mathematical expressions. A pdf or paper copy may be required for formulas and figures.

## 1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The documents should be divided into numbered sections with suitable verbal titles.

1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

## 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

## 3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in section 4.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 If possible, avoid using bold characters in formulae.

## 4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables. Use a two-level numbering system based on the section of the paper. For example, table 3.1 is the first table in section 3.

4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

## 5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

## 6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.