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

Survey Methodology

Robust variance estimators for generalized regression estimators in cluster samples

by Timothy L. Kennel and Richard Valliant

Release date: December 17, 2019

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, December 2019
Vol. 45, No. 3, pp. 427-450
Statistics Canada, Catalogue No. 12-001-X

Robust variance estimators for generalized regression estimators in cluster samples

Timothy L. Kennel and Richard Valliant[1]

Abstract

Standard linearization estimators of the variance of the general regression estimator are often too small, leading to confidence intervals that do not cover at the desired rate. Hat matrix adjustments can be used in two-stage sampling that help remedy this problem. We present theory for several new variance estimators and compare them to standard estimators in a series of simulations. The proposed estimators correct negative biases and improve confidence interval coverage rates in a variety of situations that mirror ones that are met in practice.

Key Words: Jackknife variance estimator; Hat matrix adjustment; Leverage adjustment; Superpopulation model; Two-stage sample; Sandwich variance estimator.

1 Introduction

Generalized regression (GREG) estimation is a common technique used to calibrate estimates, reduce sampling errors, and correct for nonsampling errors. Official surveys of households often use generalized regression to calibrate sample-based estimates to population controls, assure consistent estimates of demographic characteristics across surveys, and reduce nonresponse and undercoverage errors. GREG estimation is also frequently used because it draws strength from auxiliary data, resulting in smaller sampling errors than other design-based estimators.

Popular techniques used to estimate the sampling errors of calibrated estimators from complex samples either require extensive computational resources or tend to underestimate the true sampling errors, especially with small to moderate sample sizes. Two popular techniques used to estimate the sampling variance of GREG estimators are linearization and replication. Linearization estimators (Särndal, Swensson and Wretman, 1989) may not converge to the true sampling error fast enough to produce accurate results in small to moderate samples. Särndal, Swensson and Wretman (1992, page 176) remark that "For complex statistics such as an estimator of a population variance, covariance, or correlation coefficient, fairly large samples may be required before the bias is negligible." On the other hand, alternative replication techniques such as the jackknife and the bootstrap that generally produce larger variance estimates can be computationally demanding.

Leverage-adjusted sandwich estimators provide an alternative approach to estimating design-based sampling errors that also have model-based justifications. Royall and Cumberland (1978) applied this approach to develop estimators of the prediction variance of estimators of finite population totals. From a model-based framework, Long and Ervin (2000) and MacKinnon and White (1985) demonstrated how the sandwich estimator could be used for variance estimation for estimators of regression parameters even when

1. Timothy L. Kennel is an Assistant Division Chief for Statistical Methods in the Decennial Statistical Studies Division, U.S. Census Bureau. E-mail: timothy.l.kennel@census.gov; Richard Valliant is Research Professor Emeritus, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. E-mail: valliant@umich.edu.

the variance component of the working model was misspecified. Valliant (2002) took this approach to estimate the design-based variance of GREG estimators under one stage of sampling. This paper extends Valliant's work to clustered sample designs.

In Section 2, we introduce the GREG estimator and present several alternative variance estimators for it. All derivations are contained in the Appendix. In Section 3, we show how the new variance estimators perform in several simulations. In Section 4, we summarize our findings with a conclusion.

2 Theoretical results

Suppose that the population has $i = 1, 2, \ldots, M$ clusters. In cluster i there are N_i elements so that there are $N = \sum_{i=1}^{M} N_i$ elements in the population. The universe of clusters is denoted as U and the universe of elements in cluster i is U_i. An analysis variable y_{ik} is associated with element k in cluster i. The population total of y is $t_{Uy} = \sum_{i=1}^{M} \sum_{k=1}^{N_i} y_{ik}$. Each population element also has a p-vector of auxiliary variables, $\mathbf{x}_{ik}$, that can be used in estimation. A two-stage sample is selected without replacement at the first and second stages. The selection probability of cluster i is π_i, and $\pi_{k|i}$ is the conditional selection probability of element k in cluster i. The overall selection probability of element ik is $\pi_{ik} = \pi_i \pi_{k|i}$. Denote the set of sample clusters by s and the set of sample elements within cluster i by s_i. The number of sample clusters is m while the number of sample elements selected from sample cluster i is n_i. The total sample size of elements is $n = \sum_{i \in s} n_i$.

As a working model, suppose that $\mathbf{Y}_U$, the N-vector of analysis variables, follows the following linear model:

$$\begin{aligned} E_\xi(\mathbf{Y}_U) &= \mathbf{X}\boldsymbol{\beta} \\ \operatorname{cov}_\xi(\mathbf{Y}_U) &= \boldsymbol{\Psi} \end{aligned} \tag{2.1}$$

where the subscript ξ denotes expectation with respect to a model; $\mathbf{X} = [\mathbf{X}_1^\top, \mathbf{X}_2^\top, \ldots, \mathbf{X}_M^\top]^\top$ is the $N \times p$ matrix of auxiliaries with $\mathbf{X}_i$ being the $N_i \times p$ matrix of auxiliaries for the N_i elements in cluster i; and $\boldsymbol{\beta}$ is a parameter vector of length p. Elements within clusters are assumed to be correlated while elements in different clusters are independent under the model. Thus, the covariance matrix $\boldsymbol{\Psi}$ is an $N \times N$ block diagonal matrix with diagonal matrices $\boldsymbol{\Psi}_i = [\psi_{ik}]_{N_i \times N_i}$. A key feature of the variance estimators we propose is that the particular form of ψ_{ik} does not have to be known to construct variance estimators. The proposed variance estimators will be consistent regardless of the form of $\boldsymbol{\Psi}$.

Särndal et al. (1992, Chapter 8) discuss three different GREG estimators that can be used in clustered samples. These three estimators depend on the available data. We consider their case B which occurs when unit-level data are available for the complete sample and control totals are available for the population. In this case, the GREG estimator is

$$\begin{aligned} \hat{t}_y^{gr} &= \hat{t}_{y\pi} + \hat{\mathbf{B}}^\top \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi} \right) \\ &= \mathbf{g}^\top \boldsymbol{\Pi}^{-1} \mathbf{y}_s \end{aligned} \tag{2.2}$$

where $\mathbf{y}_s$ is the n-vector of y's for the sample elements, $\hat{t}_{y\pi}$ is the π-estimator of the total of the y's, $\mathbf{t}_{Ux}$ is the p-vector of population totals of the x's, $\hat{\mathbf{t}}_{x\pi}$ is the π-estimator of $\mathbf{t}_{Ux}$, and (if $\boldsymbol{\Psi}$ is known) $\hat{\mathbf{B}} = \mathbf{A}^{-1}\mathbf{X}_s^\top\boldsymbol{\Psi}_s^{-1}\boldsymbol{\Pi}^{-1}\mathbf{y}_s$ with $\mathbf{A} = \mathbf{X}_s^\top\boldsymbol{\Psi}_s^{-1}\boldsymbol{\Pi}^{-1}\mathbf{X}_s$, $\mathbf{X}_s$ the matrix of sample auxiliaries, and $\boldsymbol{\Pi} = \text{diag}[\pi_{ik}]$ $(i \in s, k \in s_i)$; $\boldsymbol{\Psi}_s$ is the part of $\boldsymbol{\Psi}$ associated with the sample elements; and $\mathbf{g}^\top = \mathbf{1}_n^\top + \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right)^\top \mathbf{A}^{-1}\mathbf{X}_s^\top\boldsymbol{\Psi}_s^{-1}$ where $\mathbf{1}_n$ is a vector of n 1's.

The component of the g-weight for sample cluster i is $\mathbf{g}_i^\top = \mathbf{1}_{n_i}^\top + \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right)^\top \mathbf{A}^{-1}\mathbf{X}_{si}^\top\boldsymbol{\Psi}_{si}^{-1}$ with $\mathbf{X}_{si}^\top = \left[\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\right]$ being the $p \times n_i$ matrix of auxiliaries for sample elements in sample cluster i, $\boldsymbol{\Psi}_{si}$ is the $n_i \times n_i$ part of $\boldsymbol{\Psi}_i$ for sample elements in sample cluster i, and $\mathbf{1}_{n_i}$ is a vector of n_i 1's. Since $\boldsymbol{\Psi}$ is generally unknown, a surrogate value $\mathbf{Q}$ may be used for $\boldsymbol{\Psi}_s^{-1}$; $\mathbf{Q} = \mathbf{I}$ is a common choice. Below, we assume that a general $\mathbf{Q}$ is used in the GREG rather than $\boldsymbol{\Psi}_s^{-1}$.

2.1 Current variance estimators

Särndal et al. (1992, Result 8.9.1) present an estimator of the design variance of $\hat{t}_y^{gr}$, which involves joint selection probabilities of clusters and elements within clusters. In the case of Poisson sampling at both stages, their estimator is

$$\upsilon_g = \sum_{i\in s}\frac{(1-\pi_i)}{\pi_i^2}\left(\hat{t}_{e,i}^{g}\right)^2 + \sum_{i\in s}\frac{1}{\pi_i}\sum_{k\in s_i}\frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2}g_{ik}^2e_{ik}^2 \tag{2.3}$$

where $\hat{t}_{e,i}^{g} = \sum_{s_i} g_{ik}e_{ik}\big/\pi_{k|i}$, g_{ik} is the k^{th} component of the $\mathbf{g}_i$ vector, and $e_{ik} = y_{ik} - \mathbf{x}_{ik}^\top\hat{\mathbf{B}}$. This estimator is computationally simpler than the general form that uses joint selection probabilities and may perform reasonably well for πps designs where the variance of estimators can be approximated by formulas that assume independence between selections.

An estimator that is appropriate if the first-stage sample is selected with replacement is

$$\upsilon_{wr} = \frac{m}{m-1}\sum_{i\in s}\left(e_{1i} - \overline{e}_1\right)^2 \tag{2.4}$$

with $e_{1i} = \sum_{k\in s_i} e_{ik}\big/\pi_{ik}$ and $\overline{e}_1 = m^{-1}\sum_{i\in s} e_{1i}$. The jackknife linearization estimator is (Yung and Rao, 1996)

$$\upsilon_{JL} = \frac{m-1}{m}\sum_{i\in s}\left(e_{2i} - \overline{e}_2\right)^2 \tag{2.5}$$

where $e_{2i} = \sum_{k\in s_i} g_{ik}e_{ik}\big/\pi_{ik}$ and $\overline{e}_2 = m^{-1}\sum_{i\in s} e_{2i}$ with g_{ik} being the k^{th} component of the $\mathbf{g}_i$ vector.

The jackknife is another popular variance estimation technique. Krewski and Rao (1981) present several asymptotically equivalent ways of writing the jackknife. The following form of the jackknife estimator is a convenient starting point for the calculations that follow:

$$\upsilon_{\text{Jack}} = \frac{m-1}{m}\sum_{i\in s}\left(\hat{t}_{y(i)}^{gr} - \hat{t}_{y(\cdot)}^{gr}\right)^2 \tag{2.6}$$

where $\hat{t}_{y(i)}^{gr}$ is the value of the GREG estimator after removing cluster i and $\hat{t}_{y(\cdot)}^{gr}$ is the average of all $\hat{t}_{y(i)}^{gr}$ estimates. Using (2.6) can be computationally demanding because m different estimates of $\hat{t}_{y(i)}^{gr}$ must be

computed. The estimators, v_{Jack}, v_{wr}, and v_{JL} are all design-consistent under the conditions in Krewski and Rao (1981) and Yung and Rao (1996). One of their key conditions is that clusters be selected with replacement. This assumption simplifies theoretical calculations but is only a convenience since the theoretical results have been shown in many empirical studies to be good predictors of estimator performance in without-replacement designs as long as the first-stage sampling fraction is small.

2.2 New variance estimators

We use the model-based framework to construct new variance estimators. First, we derive the model-based variance of $\hat{t}_y^{gr}$. Assume that model (2.1) holds and that sampling is ignorable in the sense that the probability of a unit's being in the sample given $\mathbf{Y}_U$ and $\mathbf{X}$ depends only on $\mathbf{X}$ (e.g., see discussion in Valliant, Dorfman and Royall, 2000, Section 2.6.2 and the additional references therein). Then, we construct estimators of the model variance, using hat-matrix adjustments to account for heterogeneity in the data. We evaluate the design-based properties of the new variance estimators in a simulation.

To calculate the model variance of $\hat{t}_y^{gr}$, define $\mathbf{y}_i$ as the population vector of analysis variables for cluster i, and $\mathbf{y}_{si}$ as the vector for sample elements. As shown in Appendix A.2, under model (2.1) the model-based variance of $\hat{t}_y^{gr}$ is

$$\begin{aligned}\text{var}_\xi\left(\hat{t}_y^{gr} - t_{Uy}\right) &= \sum_{i\in s}\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\boldsymbol{\Psi}_{si}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i - 2\sum_{i\in s}\left[\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\text{cov}_\xi\left(\mathbf{y}_{si},\,\mathbf{y}_i\right)\mathbf{1}_{N_i}\right] + \mathbf{1}_N^\top\boldsymbol{\Psi}\mathbf{1}_N \\ &= L_1 - 2L_2 + L_3\end{aligned}$$

where $\text{var}_\xi\left(\mathbf{y}_{si}\right) = \boldsymbol{\Psi}_{si}$, the part of $\boldsymbol{\Psi}$ associated with elements in s_i, and $\mathbf{1}_{N_i}$ and $\mathbf{1}_N$ are vectors of N_i and N 1's.

The model-based error variance of $\hat{t}_y^{gr}$ requires knowledge of $\boldsymbol{\Psi}$ for the full population. Without some strong assumptions that link the sample and nonsample covariance structures, components of $\boldsymbol{\Psi}$ associated with the nonsample cannot be estimated from the sample. However, as shown in Appendix A.2, under some reasonable conditions the orders of the terms are $L_1 = O\left(M^2/m\right)$ and $L_2 = L_3 = O\left(M\right)$ so that L_1 dominates the variance as the number of sample and population clusters increase. Thus,

$$\text{av}_\xi\left(\hat{t}_y^{gr} - t_{Uy}\right) = \sum_{i\in s}\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\boldsymbol{\Psi}_{si}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i \tag{2.7}$$

where av_ξ denotes asymptotic model variance under the assumptions in Appendix A.1. A robust estimator of the right-hand side of (2.7) can be formed even when $\boldsymbol{\Psi}_{si}$ is unknown. On the other hand, if the number of population clusters increases at the same rate as sample clusters, (i.e., $f = m/M$ converges to a non-zero constant), then L_1, L_2, and L_3 may all contribute importantly to the asymptotic variance. In this paper, we will only consider estimation of L_1.

Unless the true variance matrix of $\mathbf{y}_s$ is known, $\boldsymbol{\Psi}_i$ must be estimated. In Appendix A.3 we show that in large samples $\text{var}_\xi\left(\mathbf{e}_i\right) \approx \boldsymbol{\Psi}_i$ where $\mathbf{e}_i = \mathbf{y}_{si} - \hat{\mathbf{y}}_{si}$ with $\hat{\mathbf{y}}_{si} = \mathbf{X}_{si}\hat{\mathbf{B}}$ and $\mathbf{X}_{si}$ being the $n_i \times p$ matrix of auxiliaries for sample elements in sample cluster i. Substituting $\mathbf{e}_i\mathbf{e}_i^\top$ for $\boldsymbol{\Psi}_{si}$ in (2.7) yields the sandwich estimator

$$\upsilon_R = \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i . \tag{2.8}$$

Based on results in Appendix A.3, υ_R is approximately unbiased for $\text{av}_\xi \left(\hat{t}_y^{gr} - t_{Uy} \right)$ in large samples. This sandwich estimator is also closely related to the design-based, ultimate cluster estimator for a sample design in which clusters are selected with replacement, which is, in turn, similar to both υ_g and υ_{JL} in with replacement sampling. Consequently, υ_R has both desirable design-based and model-based properties.

In small to moderate-sized samples, υ_R will be model-biased and will often underestimate the true variance. A hat-matrix adjustment can be made as a correction. As shown in Appendix A.3,

$$\text{E}_\xi \left(\mathbf{e}_i \mathbf{e}_i^\top \right) = \text{var}_\xi \left(\mathbf{e}_i \right) = \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right) \mathbf{\Psi}_{si} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^\top + \sum_{j \neq i;\, i,\, j \in s} \mathbf{H}_{ij} \mathbf{\Psi}_{sj} \mathbf{H}_{ij}^\top \tag{2.9}$$

where $\mathbf{H}_{ij} = \mathbf{X}_{si}^\top \mathbf{A}^{-1} \mathbf{X}_{sj} \mathbf{Q}_j \mathbf{\Pi}_j^{-1}$ $(i,\ j = 1, \ldots, m)$ with $\mathbf{Q}_j$ and $\mathbf{\Pi}_j$ being the $n_j \times n_j$ parts of $\mathbf{Q}$ and $\mathbf{\Pi}$ associated with sample cluster j. As in (Li and Valliant, 2009; Valliant, 2002), the $\mathbf{H}_{ij}$ can be collected into a survey weighted hat matrix:

$$\begin{aligned} \mathbf{H} &= \mathbf{X}_s \mathbf{A}^{-1} \mathbf{X}_s^\top \mathbf{Q} \mathbf{\Pi}^{-1} \\ &= \begin{bmatrix} \mathbf{X}_{s1} \mathbf{A}^{-1} \mathbf{X}_{s1}^\top \mathbf{Q}_1 \mathbf{\Pi}_1^{-1} & \ldots & \mathbf{X}_{s1} \mathbf{A}^{-1} \mathbf{X}_{sm}^\top \mathbf{Q}_m \mathbf{\Pi}_m^{-1} \\ \vdots & \ddots & \vdots \\ \mathbf{X}_{sm} \mathbf{A}^{-1} \mathbf{X}_{s1}^\top \mathbf{Q}_1 \mathbf{\Pi}_1^{-1} & \ldots & \mathbf{X}_{sm} \mathbf{A}^{-1} \mathbf{X}_{sm}^\top \mathbf{Q}_m \mathbf{\Pi}_m^{-1} \end{bmatrix}. \end{aligned} \tag{2.10}$$

Based on the assumptions in Appendix A.1, $\mathbf{H} = O\left(m^{-1} \right)$, from which we conclude that $\text{var}_\xi \left(\mathbf{e}_i \right) \approx \mathbf{\Psi}_{si}$. The diagonal submatrices $\mathbf{H}_{ii}$ are matrix analogs to leverages in single-stage sampling. In ordinary least squares regression, the vector of predicted values can be written as $\hat{\mathbf{y}} = \mathbf{H}_{\text{OLS}} \mathbf{y}$ with $\mathbf{H}_{\text{OLS}} = \mathbf{X} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T$. Leverages are diagonals of the hat matrix, $\mathbf{H}_{\text{OLS}}$, and can be used to correct for a small sample bias in $e_i^2 = \left(y_i - \hat{y}_i \right)^2$ as an estimator of $\text{var}_\xi \left(y_i \right)$. We use the $\mathbf{H}_{ii}$ in an analogous way below.

To adjust for the fact that $\mathbf{e}_i \mathbf{e}_i^\top$ is model-biased for small to moderate samples, we make leverage-like adjustments to $\mathbf{e}_i \mathbf{e}_i^\top$. If $\mathbf{Q} = \mathbf{I}$ and the sample is self-weighting (i.e., $\mathbf{\Pi} = c\mathbf{I}$ for some $0 < c < 1$), then $\text{var}_\xi \left(\mathbf{e}_i \right) = \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right) \mathbf{\Psi}_{si}$ (see Appendix A.3). Solving for $\mathbf{\Psi}_{si}$ and substituting into (2.8) gives the variance estimator:

$$\upsilon_D = \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i \tag{2.11}$$

which, in this special case, is also approximately unbiased since $\mathbf{H}_{ii} = O\left(m^{-1} \right)$. One undesirable feature of υ_D is that it can be negative or can have negative contributions from some clusters if $\upsilon_{Di} = \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i < 0$. For such clusters, replacing υ_{Di} with $\upsilon_{Ri} = \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i$ will assure a positive variance estimator. This adjustment is used in the simulation in Section 3.

In Appendices A.4 and A.5, we show that the jackknife variance estimator can be written exactly as

$$\upsilon_{\text{Jack}} = \frac{m-1}{m}\left[\sum_{i\in s}\left(D_i - \bar{D}\right)^2 - 2\sum_{i\in s}\left(D_i - \bar{D}\right)F_i + \sum_{i\in s}F_i^2\right] \tag{2.12}$$

where

$$\begin{aligned}
F_i &= \left(G_i - \bar{G}\right) - \frac{1}{n}\left(K_i - \bar{K}\right) \\
D_i &= \mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i \\
K_i &= \left(\mathbf{1}_N^\top \mathbf{X}_U - m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_{si}\right)\left(\hat{\mathbf{B}} - \mathbf{R}_i\right);\ \bar{K} = m^{-1}\sum_{i\in s}K_i \\
G_i &= \mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\left[\mathbf{H}_{ii}\mathbf{y}_{si} - \hat{\mathbf{y}}_{si}\right];\ \bar{G} = m^{-1}\sum_{i\in s}G_i \\
\mathbf{R}_i &= \mathbf{A}^{-1}\mathbf{X}_{si}^\top \mathbf{Q}_i \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i .
\end{aligned}$$

This form of υ_{Jack} results in a significant reduction in computations since only one GREG estimate is needed, rather than m estimates. (Of course, recomputing the GREG for every jackknife replicate may still be advantageous if an elaborate nonresponse adjustment affects the size of the true variance.)

In large samples υ_{Jack} can be approximated by

$$\upsilon_{J1} = \frac{m-1}{m}\sum_{i\in s}\left(D_i - \bar{D}\right)^2 \tag{2.13}$$

or by

$$\begin{aligned}
\upsilon_{J2} &= \frac{m-1}{m}\sum_{i\in s}D_i^2 \\
&= \frac{m-1}{m}\sum_{i\in s}\mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\mathbf{e}_i^\top\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i .
\end{aligned} \tag{2.14}$$

The estimators, υ_{J1} and υ_{J2} are clustered versions of the single-stage approximations to the jackknife in Valliant (2002, equations (3.5), (3.6)).

As sketched in Appendix A.6, $\upsilon_{\text{Jack}}, \upsilon_{JL}, \upsilon_{J1}, \upsilon_{J2}, \upsilon_D,$ and υ_R are all asymptotically equivalent as $m \to \infty$. Since υ_{Jack} and υ_{JL} are design-consistent, the alternative estimators above can be expected to perform well over repeated samples when the size of the first-stage sample is large, and when model (2.1) is approximately correct. One caveat is that the sampling fraction of clusters must be small so that estimators made from a without-replacement, first-stage sample will perform as if the sample had been selected with-replacement.

None of these sandwich-like estimators includes finite population correction factors. Thus, they may tend to overestimate the sampling variance when a large proportion of the sample clusters is selected. To account for this, we can further adjust all of the variance estimators in an ad hoc fashion by multiplying the variance estimators by a finite population correction factor, denoted f_{pc}, as developed by Kott (1988). This results in the following adjusted estimators:

$$
\begin{aligned}
\upsilon_R^* &= f_{pc} \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i \\
\upsilon_D^* &= f_{pc} \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i \\
\upsilon_{\text{Jack}}^* &= f_{pc} \frac{m}{m-1} \left[\sum_{i \in s} \left(D_i - \bar{D} \right)^2 - 2 \sum_{i \in s} \left(D_i - \bar{D} \right) F_i + \sum_{i \in s} F_i^2 \right] \\
\upsilon_{J1}^* &= f_{pc} \frac{m}{m-1} \sum_{i \in s} \left(D_i - \bar{D} \right)^2 \\
\upsilon_{J2}^* &= f_{pc} \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{\Pi}_i^{-1} \mathbf{g}_i .
\end{aligned}
$$

When a simple random sample is selected in the first stage, $f_{pc} = 1 - m/M$. According to Kott (1988), an appropriate correction when the first stage is selected with varying probabilities is $f_{pc} = 1 - m\sum_{i=1}^{M} p_i^2$ where p_i is the single draw probability for cluster i, i.e., the probability that cluster i would be selected in a sample of size 1.

3 Simulation

We performed a series of simulation studies to test the performance of the new variance estimators in different populations. In each simulated sample, we computed the quantities listed in Table 3.1. To evaluate the variance estimators, we calculated the average of the variance estimates, compared those averages to the empirical mean square error, and computed coverage probabilities of confidence intervals based on the different variance estimates. Table 3.2 summarizes the sample designs for the 18 simulation studies. The column called Label gives the headings used in later tables. The sample designs are used in three populations described below.

Table 3.1
Statistics of interest for clustered GREG variance simulation

Statistic	Description
$\hat{t}_y^{\pi}$	Estimated total from the Horvitz-Thompson Estimator
$\hat{t}_y^{gr}$	Estimated total from the GREG
υ_E	Empirical variance
υ_g	Design-based variance estimator that assumes Poisson sampling at both stages from Särndal et al. (1992) in (2.3)
υ_{wr}	With-replacement variance estimator in (2.4)
υ_{JL}	Jackknife linearization variance estimator from Yung and Rao (1996) in (2.5)
υ_R	Sandwich estimator in (2.8)
υ_D	First hat-matrix adjusted sandwich estimator in (2.11)
υ_{Jack}	Jackknife variance estimator in (2.6)
υ_{J1}	First approximation to the jackknife variance estimator in (2.13)
υ_{J2}	Second approximation to the jackknife variance estimator in (2.14)
υ_R^*	Sandwich estimator with a finite population adjustment
υ_D^*	First hat-matrix adjusted sandwich estimator with a finite population correction
υ_{Jack}^*	Jackknife variance estimator with a finite population correction
υ_{J1}^*	First approximation to jackknife with a finite population correction
υ_{J2}^*	Second approximation to jackknife with a finite population adjustment

3.1 Data

We conducted simulations on three populations to assess the design-based performance of the variance estimators under a variety of situations. In the first population, we investigated the performance of the variance estimators when the first-stage sampling fraction was large and the sample size was moderate. The focus of the second simulation study was on the performance of the variance estimators under a relatively messy dataset and a small first-stage sample size. The final simulation study shows the performance of the variance estimators in large samples.

Table 3.2
Simulation designs for three populations

	Label	Population	First stage sample	*m*	Second stage sample	No. of samples
1	srs fixed	Third Grade	srswor	25	$n_i = 5$	1,000
2	srs fixed	Third Grade	srswor	50	$n_i = 5$	1,000
3	srs epsem	Third Grade	srswor	25	$f_i = \frac{675}{2,427}$	1,000
4	srs epsem	Third Grade	srswor	50	$f_i = \frac{675}{2,427}$	1,000
5	pps epsem	Third Grade	ppswor	25	$n_i = 5$	1,000
6	pps epsem	Third Grade	ppswor	50	$n_i = 5$	1,000
7	srs fixed	ACS	srswor	3	$n_i = 9$	5,000
8	srs fixed	ACS	srswor	15	$n_i = 9$	5,000
9	srs epsem	ACS	srswor	3	$f_i = \frac{30,430}{194,329}$	5,000
10	srs epsem	ACS	srswor	15	$f_i = \frac{30,430}{194,329}$	5,000
11	pps epsem	ACS	ppswor	3	$n_i = 9$	5,000
12	pps epsem	ACS	ppswor	15	$n_i = 9$	5,000
13	srs fixed	Simulated	srswor	300	$n_i = 2$	1,000
14	srs fixed	Simulated	srswor	1,500	$n_i = 2$	100
15	srs epsem	Simulated	srswor	300	$f_i = \frac{60,000}{195,164}$	1,000
16	srs epsem	Simulated	srswor	1,500	$f_i = \frac{60,000}{195,164}$	100
17	pps epsem	Simulated	ppswor	300	$n_i = 3$	1,000
18	pps epsem	Simulated	ppswor	1,500	$n_i = 3$	100

3.1.1 Third grade population

The first simulation study used the Third Grade population from Appendix B.6 of Valliant et al. (2000). This dataset contained the mathematics achievement scores for 2,427 third graders in 135 schools. The relatively small number of schools in this population and the fairly constant number of students in each school made it ideal for studying samples with large sampling fractions.

We used GREG to estimate the average mathematics achievement score for third graders. Altogether, we selected 1,000 samples in each of six sample designs listed in Table 3.2. In the first sample design, we selected 1,000 simple random samples without replacement (srswor) of 25 schools. Within each sampled school, we selected exactly five students via srswor. Because the number of students in each school varied

from school to school, this sample design resulted in different unconditional probabilities of selection, but a fixed sample size of 125 students. The second sample design was similar to the first, except we selected 50 schools. Selecting 50 of the 135 schools resulted in a large first-stage sampling fraction of 0.37, necessitating a finite population correction factor. Both the samples of $m = 25$ and 50 might be considered to be of "moderate" size.

In the third sample design, we selected 1,000 simple random samples of 25 schools without replacement. Within each sampled school, we selected students at a constant rate of $\frac{675}{2,427}$, yielding 1,000 samples with random sizes centered around 125 students. The result of this design was that each student had the same unconditional probability of selection. The fourth sample design was similar to the third, except we selected 50 schools. The sample sizes were also random under this design, with an average of 250 students. Since the third and fourth sample designs resulted in every unit getting the same chance of selection, these sample designs are labeled srs epsem (equal probability selection mechanism) in subsequent tables.

In the fifth design, we selected 1,000 samples of 25 schools with probabilities proportional to the number of students in each school. Within each sampled school, we selected exactly five students, yielding 1,000 samples with exactly 125 students each. The sixth sample design was similar to the fifth, except we selected 50 schools. We selected 1,000 samples of size 250 students using this design. The fifth and sixth designs are epsem. Like the second and fourth sample designs, this sample design also had a large sampling fraction and warranted the need for a finite population correction factor to adjust the variance estimators.

From each sample, we estimated the average achievement scores for the finite population using a GREG estimator and assuming that the number of students in the population was known. The assisting model was meant to replicate the clustered linear regression model in Section 9.6 of Valliant et al. (2000). The eleven explanatory variables used to model each student's math achievement score were: an intercept, sex (male or female), ethnicity (White/Asian, Black, Native American/Other, or Hispanic), language spoken at home is the same as the test (Always, Sometimes/Never), type of community (Outskirts of a town or city, Village/City), and school enrollment. The total mathematics achievement estimated with the GREG estimator was divided by the number of students in the population, 2,427, to get the average achievement score. The average achievement score for the population was 477.7. For the full population, the R-squared for the student-level linear model was 0.9735, indicating a very strong linear relationship.

3.1.2 American Community Survey population

The second simulation study used Census 2000 Summary File 3 data and American Community Survey (ACS) 2005 - 2009 Summary File data. The goal was to estimate the total number of housing units in the U.S. state of Alabama as reported in the ACS Summary File. Block group counts from Census 2000 were used as covariates in the assisting model.

To create the population, first all block group data were extracted from the ACS Summary File and the Census 2000 Summary File 3. Then, the two files were merged at the block group level. Block groups with 1,000 or more housing units in Census 2000 were removed because such large block groups had different characteristics than the majority of blocks. In many sampling designs such large units would be placed in a separate, certainty stratum and not contribute to the variance of estimates. Also, block groups with extreme

growth in the total number of housing units were also removed. Specifically block groups that had gained more than 10 units over twice the 2000 census count were removed.

Clusters were defined as counties and block groups were treated as units. Treating block group as a unit is motivated by the common task of selecting a sample of blocks, listing them, and then using the listings to estimate the total number of housing units in the finite population.

Clusters with fewer than 10 block groups or more than 120 block groups in them were removed from the frame of clusters. Overall, there were 61 clusters (counties) containing a total of 2,051 block groups and 1,109,499 housing units in the edited dataset. Altogether, six counties and 1,278 block groups containing 1,030,471 housing units were removed from the Alabama file.

Figure 3.1 shows two scatterplots. The first plot shows the total number of housing units in the block group as reported on the ACS summary file as a function of the 2000 Census housing unit count. Each point represents one of the 2,051 block groups in the finite population. The diagonal line is a nonparametric smoother, indicating a strong relationship between the two variables. The plot also shows some evidence of heteroscedasticity because the points appear to fan out as the 2000 census count increases. The second plot shows the residuals obtained by regressing the 2000 census housing unit count on the ACS housing unit count using ordinary least squares (OLS) plotted versus the ACS housing unit count. As the number of housing units reported on the ACS file increases, the model predictions appear to seriously underestimate the true number of housing units. This suggests some degree of nonlinearity in the mean function. In addition, there is noticeable heteroscedasticity in variance.

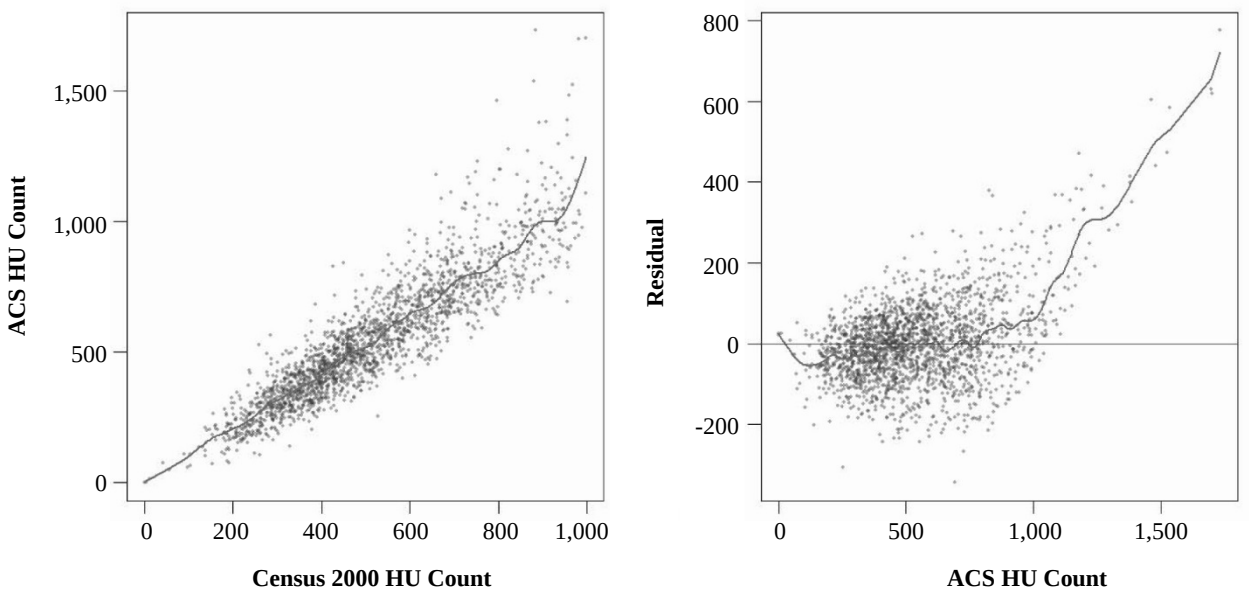


Figure 3.1 Scatter plot and residual plot for ACS population. Gray lines are nonparametric smoothers.

As in the first simulation study, we tested six different sample designs. We selected 5,000 samples in each of six different selection mechanisms listed in Table 3.2. In the first sample design, we selected 5,000 simple random samples of 3 clusters without replacement. In large national surveys, it is not uncommon to

select a small number of primary sampling units in each stratum. In this case, we treat Alabama as if it were a single design stratum and its 61 counties as clusters. Three counties within that stratum were sampled. Within each cluster, we selected nine block groups using srswor. The second design was similar with 15 clusters and 9 block groups per cluster. The first two sample designs resulted in highly variable weights. The other designs (rows 9-12) were parallel to those in rows 3-6 for the Third Grade population. The sample sizes of $m = 3$ and 15 are small so that theoretical, large sample properties are less likely to hold.

From each sample, we estimated the total number of housing units in the finite population using a GREG estimator. The assisting model included an intercept and the Census 2000 count of housing units; the heteroscedasticity noted above was not accounted for in the GREG. For the full population, the R-squared was 0.819, again indicating a strong linear relationship.

3.1.3 Simulated population

A population was created with a large number of clusters to assess the asymptotic characteristics of the variance estimators. Generated using a classic linear model, a total of 30,000 clusters were created, each with a random number of units. The number of units in each cluster was determined by adding three to a uniform random integer between 0 and 7. This created clusters ranging in size from 3 to 10 units. Altogether, the population contained 195,164 units within 30,000 clusters. For each unit, a positive covariate was created as $x_k \sim 1{,}000 \exp N(0, 1)$ where $N(0, 1)$ is a normal random variate with mean of 0 and standard deviation of 1. A random response was created such that $y_k \sim N\left(1{,}000 + 2x_k, \frac{x_k}{2}\right)$. Figure 3.2 shows scatter plots of the relationship between x_k and y_k for the finite population.

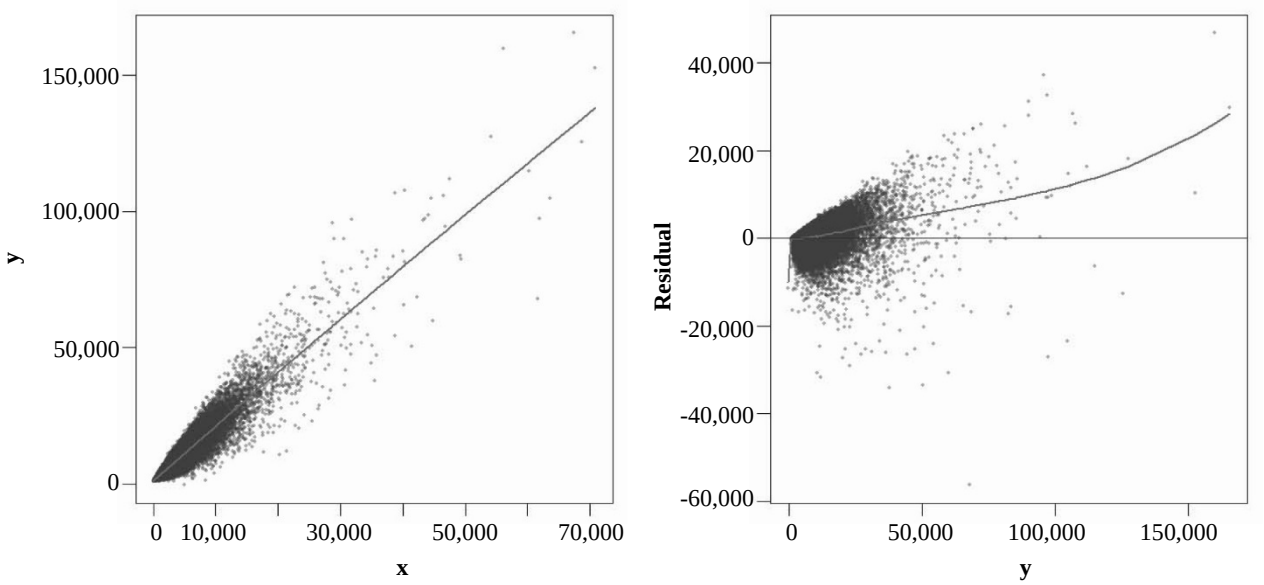


Figure 3.2 Scatter plot and residual for simulated population. Gray lines are nonparametric smoothers.

We selected samples using the six different probability selection mechanisms listed in rows 13-18 of Table 3.2. The types of sample designs are parallel to those used for the Third Grade and ACS populations.

In designs 14, 16, and 18, we selected 100 simple random samples of 1,500 clusters without replacement. We only selected 100 samples due to the excessive amount of computer time it took to select and process each sample. The sample sizes of $m = 300$ and 1,500 are large so that theoretical, large sample properties should hold.

From each sample, we estimated the total of the response using a GREG estimator. The true finite population total was 839,149,969. The assisting model included an intercept and x with $\mathbf{Q} = \mathbf{I}$. For the full population, the R-squared was 0.953, indicating a very strong linear relationship. Figure 3.2 shows a scatter plot of the population as well as a residual plot based on an OLS regression of x_k on y_k for the full population. There is clear evidence of heteroscedasticity of errors.

3.2 Results

We explored the bias, variability, and confidence interval coverage of the new and existing variance estimators. We only show tables for some of the simulations to conserve space. Table 3.3 shows the means of the π-estimator and the GREG estimator as well as the ratios of the average values of the variance estimators to the empirical mse's for all populations and sample size combinations across all simulations. Both the π-estimator and the GREG estimator are approximately unbiased; however, the GREG estimator is much more efficient.

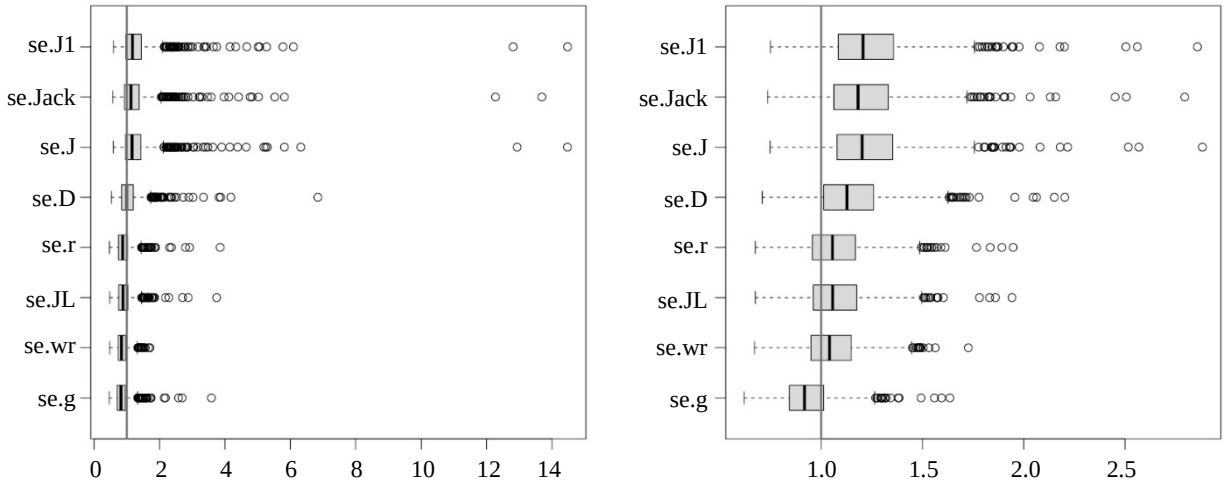
Table 3.3
Simulation Results for estimates for means and variance estimators for three populations and six sample designs in each population. Values in rows for variance estimators are ratios of mean estimated variance to empirical mse of the GREG. See Table 3.1 for descriptions of the variance estimators

Estimator	srs fixed		srs epsem		pps epsem	
			Third Grade Population			
	$m = 25$	$m = 50$	$m = 25$	$m = 50$	$m = 25$	$m = 50$
Average $\hat{t}_y^{\pi} / N$	477.23	477.11	476.29	476.85	477.31	477.75
mse $\hat{t}_y^{\pi} / N$	663.12	264.75	2,013.90	981.54	142.93	53.17
Average $\hat{t}_y^{g} / N$	474.27	476.37	476.95	477.24	477.50	477.85
mse $\hat{t}_y^{g} / N$	218.96	66.66	114.08	50.10	121.57	41.32
$v_g / \text{mse}(\hat{t}_y^{g})$	0.76	0.87	0.73	0.82	0.66	0.91
$v_{wr} / \text{mse}(\hat{t}_y^{g})$	0.75	1.11	0.79	1.06	0.73	1.19
$v_{JL} / \text{mse}(\hat{t}_y^{g})$	0.88	1.16	0.85	1.10	0.78	1.24
$v_R / \text{mse}(\hat{t}_y^{g})$	0.87	1.15	0.82	1.08	0.74	1.22
$v_D / \text{mse}(\hat{t}_y^{g})$	1.26	1.32	1.09	1.25	0.95	1.36
$v_{J2} / \text{mse}(\hat{t}_y^{g})$	2.22	1.54	1.50	1.46	1.23	1.54
$v_{\text{Jack}} / \text{mse}(\hat{t}_y^{g})$	2.03	1.49	1.44	1.43	1.19	1.51
$v_{J1} / \text{mse}(\hat{t}_y^{g})$	2.22	1.55	1.56	1.49	1.28	1.57
$v_R^* / \text{mse}(\hat{t}_y^{g})$	0.71	0.73	0.67	0.68	0.60	0.74
$v_D^* / \text{mse}(\hat{t}_y^{g})$	1.02	0.83	0.88	0.79	0.76	0.83
$v_{J2}^* / \text{mse}(\hat{t}_y^{g})$	1.81	0.97	1.22	0.92	0.99	0.93
$v_{\text{Jack}}^* / \text{mse}(\hat{t}_y^{g})$	1.66	0.94	1.17	0.90	0.95	0.92
$v_{J1}^* / \text{mse}(\hat{t}_y^{g})$	1.81	0.98	1.27	0.94	1.03	0.95

Table 3.3 (continued)
Simulation Results for estimates for means and variance estimators for three populations and six sample designs in each population. Values in rows for variance estimators are ratios of mean estimated variance to empirical mse of the GREG. See Table 3.1 for descriptions of the variance estimators

Estimator	srs fixed		srs epsem		pps epsem	
	ACS Population (numbers in thousands)					
	$m = 3$	$m = 15$	$m = 3$	$m = 15$	$m = 3$	$m = 15$
Average $\hat{t}_y^{\pi} / N$	1,119.13	1,108.23	1,112.89	1,113.89	1,111.48	1,109.02
mse $\hat{t}_y^{\pi} / N$	181,329.24	27,650.01	201,618.77	32,926.98	15,991.69	2,619.32
Average $\hat{t}_y^{g} / N$	1,081.68	1,103.34	1,104.45	1,108.45	1,106.36	1,108.46
mse $\hat{t}_y^{g} / N$	11,220.86	921.82	2,111.84	408.19	1,874.39	352.65
$\upsilon_g / \text{mse}(\hat{t}_y^{g})$	2.70	0.90	0.44	0.83	0.53	0.92
$\upsilon_{wr} / \text{mse}(\hat{t}_y^{g})$	1.17	0.98	0.68	1.03	0.87	1.14
$\upsilon_{JL} / \text{mse}(\hat{t}_y^{g})$	2.18	0.91	0.65	0.99	0.79	1.11
$\upsilon_R / \text{mse}(\hat{t}_y^{g})$	2.80	1.00	0.43	0.92	0.53	1.03
$\upsilon_D / \text{mse}(\hat{t}_y^{g})$	6.09	1.32	0.84	1.08	0.89	1.15
$\upsilon_{J2} / \text{mse}(\hat{t}_y^{g})$	17,191.52	1.85	2.36	1.27	1.64	1.29
$\upsilon_{\text{Jack}} / \text{mse}(\hat{t}_y^{g})$	4,678.25	1.47	1.37	1.19	1.05	1.21
$\upsilon_{J1} / \text{mse}(\hat{t}_y^{g})$	17,190.86	1.72	3.07	1.36	2.35	1.38
$\upsilon_R^* / \text{mse}(\hat{t}_y^{g})$	2.66	0.76	0.41	0.70	0.49	0.68
$\upsilon_D^* / \text{mse}(\hat{t}_y^{g})$	5.79	0.99	0.80	0.82	0.83	0.76
$\upsilon_{J2}^* / \text{mse}(\hat{t}_y^{g})$	16,346.03	1.40	2.25	0.96	1.52	0.85
$\upsilon_{\text{Jack}}^* / \text{mse}(\hat{t}_y^{g})$	4,448.17	1.11	1.30	0.90	0.97	0.80
$\upsilon_{J1}^* / \text{mse}(\hat{t}_y^{g})$	16,345.41	1.30	2.92	1.03	2.19	0.91
	Simulated Population (numbers in millions)					
	$m = 300$	$m = 1{,}500$	$m = 300$	$m = 1{,}500$	$m = 300$	$m = 1{,}500$
Average $\hat{t}_y^{\pi} / N$	838.91	838.71	838.13	843.13	838.74	839.06
mse $\hat{t}_y^{\pi} / N$	1,588.43	250.20	2,303.19	563.77	1,218.73	253.13
Average $\hat{t}_y^{g} / N$	838.57	839.10	838.81	840.01	839.39	839.08
mse $\hat{t}_y^{g} / N$	156.29	23.07	117.18	19.63	105.64	25.24
$\upsilon_g / \text{mse}(\hat{t}_y^{g})$	0.91	1.11	0.91	1.13	1.01	0.89
$\upsilon_{wr} / \text{mse}(\hat{t}_y^{g})$	0.94	1.13	0.91	1.17	1.01	0.90
$\upsilon_{JL} / \text{mse}(\hat{t}_y^{g})$	0.91	1.13	0.92	1.15	1.02	0.90
$\upsilon_R / \text{mse}(\hat{t}_y^{g})$	0.91	1.13	0.92	1.14	1.02	0.90
$\upsilon_D / \text{mse}(\hat{t}_y^{g})$	1.03	1.15	0.96	1.16	1.07	0.91
$\upsilon_{J2} / \text{mse}(\hat{t}_y^{g})$	1.50	1.17	1.03	1.18	1.13	0.93
$\upsilon_{\text{Jack}} / \text{mse}(\hat{t}_y^{g})$	1.48	1.17	1.03	1.18	1.12	0.93
$\upsilon_{J1} / \text{mse}(\hat{t}_y^{g})$	1.50	1.17	1.03	1.18	1.13	0.93
$\upsilon_R^* / \text{mse}(\hat{t}_y^{g})$	0.90	1.07	0.91	1.09	1.01	0.85
$\upsilon_D^* / \text{mse}(\hat{t}_y^{g})$	1.02	1.09	0.96	1.11	1.05	0.86
$\upsilon_{J2}^* / \text{mse}(\hat{t}_y^{g})$	1.48	1.11	1.02	1.12	1.12	0.88
$\upsilon_{\text{Jack}}^* / \text{mse}(\hat{t}_y^{g})$	1.47	1.11	1.01	1.12	1.11	0.88
$\upsilon_{J1}^* / \text{mse}(\hat{t}_y^{g})$	1.48	1.11	1.02	1.13	1.12	0.88

The performance of the variance estimators depends on the sample design and the population. Some of the estimates in Table 3.3 from the ACS population with the simple random sample of 3 clusters and 9 units in each cluster stand out as being extremely poor. The inverses of the probabilities of selection vary quite a bit for this sample design. The variability of these weights, coupled with some extreme observations in the population, causes instability for some of the variance estimators. Namely, $\upsilon_{J2}, \upsilon_{\text{Jack}}, \upsilon_{J1}, \upsilon^*_{J2}, \upsilon^*_{\text{Jack}}, \upsilon^*_{J1}$ are extreme overestimates on average. All six of these estimators contain explicit or implicit hat matrix adjustments which can be quite large and seriously inflate the variance estimators when coupled with large sampling weights. On the other hand, υ_D, which also has a hat matrix adjustment, performs reasonably well for all populations and sample sizes. Noteworthy is the result that υ_D is much less of an overestimate for the mse in the combination (ACS, srs fixed, $m = 3$, $n_i = 9$) whereas other hat-matrix adjusted estimators were extreme overestimates. The estimators, $\upsilon_g, \upsilon_{wr}$, and to a lesser extent, υ_R and υ_{JL}, tend to be underestimates at the smaller sample sizes in the Third Grade and ACS populations and for all sample designs in those populations, but the problem diminishes for the larger sample sizes.



(a) Samples of size $m = 25$ (b) Samples of size $m = 50$

Figure 3.3 Boxplots of ratios of standard error estimates to the empirical standard errors for 1,000 SRS samples from Third Grade population. Vertical reference lines at 1.

The boxplots in Figure 3.3 show the variability of the estimators more clearly for srs's of size $m = 25$ and 50 from the Third Grade population. The boxplots depict the estimated standard errors (SEs) as a fraction of the empirical SE for the samples in each simulation. A ratio of 1 means that the estimated variance was equal to the empirical variance. Some samples yield large SE estimates, even though the majority of samples are much closer to the empirical variance. The degree of overestimation and the incidence of extreme values decreases substantially with the larger sample size as is evident by comparing the figures. The hat-matrix adjusted estimators also tend to somewhat overestimate the true variance, as evinced by the

boxes that are shifted above the reference lines drawn at 1. This can be an advantage for confidence interval coverage.

Table 3.4 shows the six-number summaries of the ratios of the SE estimates, $\sqrt{v}$, to the square root of the empirical variance, $\sqrt{v_E}$, for the Third Grade population for four of the sample designs. As indicated by the median value of the ratios for υ_{J2}, υ_{Jack}, υ_{J1}, υ^*_{J2}, υ^*_{Jack}, and υ^*_{J1}, they are generally centered near the empirical SEs, but can have extremely large values in some samples that affect their averages. (The problem of outlying values is even more severe in the ACS population; details are not shown here.) The estimators that are least affected by extremes are υ_g, υ_{wr}, υ_{JL}, υ_R, υ_D, υ^*_R, and υ^*_D. However, the estimators that incorporate *fpc*'s often are underestimates except in the case of srs and $m = 25$.

Table 3.4
Six-number summaries for alternative standard error estimators for Third Grade population in four sample designs. v_E is empirical variance across simulated samples. See Table 3.1 for descriptions of the variance estimators

	$\sqrt{v}$	Distribution of $\sqrt{v}/\sqrt{\upsilon_E}$					
		Min	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
srs $m = 25$	$\sqrt{\upsilon_g}$	0.46	0.71	0.82	0.86	0.96	3.59
	$\sqrt{\upsilon_{wr}}$	0.48	0.73	0.84	0.87	0.97	1.71
	$\sqrt{\upsilon_{JL}}$	0.48	0.75	0.88	0.92	1.03	3.75
	$\sqrt{\upsilon_R}$	0.47	0.74	0.87	0.92	1.02	3.85
	$\sqrt{\upsilon_D}$	0.53	0.84	1.00	1.08	1.20	6.84
	$\sqrt{\upsilon_{J2}}$	0.59	0.96	1.16	1.31	1.43	14.47
	$\sqrt{\upsilon_{\text{Jack}}}$	0.57	0.93	1.13	1.26	1.38	13.69
	$\sqrt{\upsilon_{J1}}$	0.59	0.97	1.17	1.32	1.44	14.48
	$\sqrt{\upsilon^*_R}$	0.42	0.67	0.79	0.83	0.92	3.48
	$\sqrt{\upsilon^*_D}$	0.48	0.76	0.90	0.97	1.08	6.17
	$\sqrt{\upsilon^*_{J2}}$	0.53	0.87	1.05	1.18	1.29	13.06
	$\sqrt{\upsilon^*_{\text{Jack}}}$	0.52	0.84	1.02	1.14	1.25	12.35
	$\sqrt{\upsilon^*_{J1}}$	0.54	0.88	1.06	1.19	1.30	13.07
srs $m = 50$	$\sqrt{\upsilon_g}$	0.62	0.84	0.92	0.94	1.01	1.64
	$\sqrt{\upsilon_{wr}}$	0.67	0.95	1.04	1.06	1.15	1.73
	$\sqrt{\upsilon_{JL}}$	0.68	0.96	1.06	1.08	1.18	1.94
	$\sqrt{\upsilon_R}$	0.68	0.96	1.06	1.07	1.17	1.95
	$\sqrt{\upsilon_D}$	0.71	1.01	1.13	1.15	1.26	2.20
	$\sqrt{\upsilon_{J2}}$	0.75	1.08	1.20	1.24	1.35	2.88
	$\sqrt{\upsilon_{\text{Jack}}}$	0.74	1.06	1.18	1.22	1.33	2.79
	$\sqrt{\upsilon_{J1}}$	0.75	1.09	1.21	1.24	1.36	2.86
	$\sqrt{\upsilon^*_R}$	0.54	0.76	0.84	0.85	0.93	1.55
	$\sqrt{\upsilon^*_D}$	0.56	0.80	0.89	0.91	1.00	1.75
	$\sqrt{\upsilon^*_{J2}}$	0.59	0.86	0.95	0.98	1.07	2.29
	$\sqrt{\upsilon^*_{\text{Jack}}}$	0.58	0.84	0.94	0.97	1.06	2.22
	$\sqrt{\upsilon^*_{J1}}$	0.60	0.86	0.96	0.99	1.08	2.27

Table 3.4 (continued)
Six-number summaries for alternative standard error estimators for Third Grade population in four sample designs. v_E is empirical variance across simulated samples. See Table 3.1 for descriptions of the variance estimators

	$\sqrt{v}$	Distribution of $\sqrt{v}/\sqrt{v_E}$					
		Min	1st Qu.	Median	Mean	3rd Qu.	Max
pps $m = 25$	$\sqrt{v_g}$	0.48	0.71	0.79	0.80	0.88	1.33
	$\sqrt{v_{wr}}$	0.51	0.76	0.84	0.84	0.92	1.30
	$\sqrt{v_{JL}}$	0.50	0.76	0.86	0.87	0.96	1.46
	$\sqrt{v_R}$	0.49	0.75	0.84	0.85	0.94	1.43
	$\sqrt{v_D}$	0.53	0.83	0.94	0.96	1.06	1.66
	$\sqrt{v_{J2}}$	0.59	0.94	1.06	1.09	1.21	2.15
	$\sqrt{v_{\text{Jack}}}$	0.57	0.92	1.04	1.07	1.18	2.10
	$\sqrt{v_{J1}}$	0.60	0.96	1.08	1.11	1.23	2.19
	$\sqrt{v_R^*}$	0.43	0.67	0.76	0.76	0.84	1.30
	$\sqrt{v_D^*}$	0.47	0.75	0.84	0.86	0.95	1.51
	$\sqrt{v_{J2}^*}$	0.52	0.84	0.95	0.98	1.08	1.90
	$\sqrt{v_{\text{Jack}}^*}$	0.51	0.82	0.93	0.96	1.06	1.86
	$\sqrt{v_{J1}^*}$	0.53	0.86	0.97	1.00	1.10	1.93
pps $m = 50$	$\sqrt{v_g}$	0.72	0.88	0.95	0.95	1.01	1.28
	$\sqrt{v_{wr}}$	0.78	1.00	1.09	1.09	1.16	1.47
	$\sqrt{v_{JL}}$	0.81	1.01	1.11	1.11	1.19	1.52
	$\sqrt{v_R}$	0.80	1.00	1.09	1.09	1.18	1.50
	$\sqrt{v_D}$	0.84	1.06	1.15	1.16	1.25	1.64
	$\sqrt{v_{J2}}$	0.88	1.11	1.22	1.23	1.33	1.83
	$\sqrt{v_{\text{Jack}}}$	0.88	1.10	1.21	1.22	1.31	1.81
	$\sqrt{v_{J1}}$	0.89	1.13	1.23	1.24	1.34	1.85
	$\sqrt{v_R^*}$	0.62	0.78	0.85	0.85	0.92	1.16
	$\sqrt{v_D^*}$	0.65	0.82	0.90	0.90	0.97	1.28
	$\sqrt{v_{J2}^*}$	0.68	0.87	0.95	0.96	1.03	1.43
	$\sqrt{v_{\text{Jack}}^*}$	0.67	0.86	0.94	0.95	1.02	1.42
	$\sqrt{v_{J1}^*}$	0.69	0.88	0.96	0.97	1.04	1.44

Lastly, Table 3.5 shows the 95% confidence interval coverage for all of the estimators based on t-distributions. That is, we computed, $\left[\hat{t}_y^{gr} - t_{0.975,\, m-1}\sqrt{v},\ \hat{t}_y^{gr} + t_{0.975,\, m-1}\sqrt{v}\ \right]$ where $t_{0.975,\, m-1}$ is the 97.5th percentile from a t-distribution with $m-1$ degrees of freedom. We then noted how often the true value fell below, above, and inside this range. In addition to the new and old estimators, Table 3.5 also shows the confidence interval coverage attained when the empirical variance, v_E, was used to form the confidence intervals. Ideally, the population total should be within the estimated 95% confidence interval for 95% of the samples. The true total should be below the 95% confidence bounds for 2.5% of the samples and above the confidence bounds for the same percentage of samples.

The jackknife-based estimators, v_D^*, v_{Jack}^*, and v_{J2}, cover at higher rates than the other variance estimators because they are larger. In small samples, jackknife-based estimators cover above the nominal level. The traditional variance estimators, v_g, v_{wr}, and v_{JL} under-covered in a number of cases, although their coverage was almost always higher than 90%. Note that v_g is generally an improvement over v_R due to the hat-matrix adjustment that makes v_D larger.

The variance estimators that incorporate hat matrix adjustments (v_D, v_{J2}, v_{Jack}, and v_R^*) generally increase CI coverage rates compared to the other choices. This advantage was especially noticeable for the ACS population where, for example, v_{wr} covers in less than 90% of samples in the combinations, (v_{Jack}^*, $m = 3$), (srs epsem, $m = 3$), and (srs epsem, $m = 15$). Although, in principal, an *fpc* would seem useful in some of the population and sample size combinations, CIs based on the variance estimators with *fpc*'s cover at lower rates than their counterparts without the *fpc*'s. For example, in ACS (srs epsem, $m = 15$) the coverage rates for v_R^*, v_D^*, v_{J2}^*, v_{Jack}^*, and v_{J1}^* range from 86.1 to 90.6% while the rates for the versions without *fpc*'s range from 90.2 to 93.4%.

Table 3.5
Coverage of 95% confidence intervals for population totals based on *t*-distributions and alternative variance estimators. See Table 3.1 for descriptions of the variance estimators

Variance est.	Third Grade			ACS			Simulation			Third Grade			ACS			Simulation		
	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper
	srs $m = 25$			srs $m = 3$			srs $m = 300$			srs $m = 50$			srs $m = 15$			srs $m = 1{,}500$		
v_E	2.9	95.6	1.5	0.7	99.3	0	2.7	95.0	2.3	3.4	95.1	1.5	3.3	95.8	1.0	1.0	96.0	3.0
v_a	7.4	90.7	1.9	2.4	97.3	0.4	4.3	93.5	2.2	5.9	92.8	1.3	6.6	92.3	1.0	1.0	95.0	4.0
v_{wr}	7.0	90.5	2.5	9.2	88.8	2.0	3.9	92.8	3.3	4.1	95.0	0.9	7.5	91.0	1.5	1.0	96.0	3.0
v_{JL}	5.5	93.2	1.3	6.5	92.1	1.4	4.4	93.4	2.2	3.3	96.1	0.6	7.2	91.4	1.4	1.0	95.0	4.0
v_R	5.9	92.7	1.4	3.1	96.3	0.6	4.3	93.5	2.2	3.4	96.0	0.6	6.5	92.5	1.0	1.0	95.0	4.0
v_D	3.8	95.4	0.8	1.6	98.0	0.4	3.7	94.2	2.1	2.4	97.1	0.5	5.1	94.3	0.6	1.0	95.0	4.0
v_{J2}	1.7	98.0	0.3	0.6	99.3	0.1	3.6	94.4	2.0	2.0	97.7	0.3	3.9	95.7	0.4	1.0	95.0	4.0
v_{Jack}	2.1	97.6	0.3	3.2	95.9	0.8	3.6	94.4	2.0	2.0	97.7	0.3	5.6	93.7	0.7	1.0	95.0	4.0
v_{J1}	1.6	98.1	0.3	1.6	98.0	0.3	3.6	94.4	2.0	2.0	97.7	0.3	4.5	95.0	0.5	1.0	95.0	4.0
v_R^*	8.6	89.4	2.0	3.4	96.0	0.7	4.4	93.4	2.2	7.8	89.8	2.4	9.5	88.5	2.0	1.0	95.0	4.0
v_D^*	5.5	93.3	1.2	1.6	98.0	0.4	3.8	94.1	2.1	6.4	92.2	1.4	7.5	91.1	1.4	1.0	95.0	4.0
v_{J2}^*	2.9	96.6	0.5	0.6	99.3	0.1	3.6	94.4	2.0	5.2	93.8	1	5.8	93.3	0.8	1.0	95.0	4.0
v_{Jack}^*	3.7	95.7	0.6	3.4	95.7	0.9	3.6	94.4	2.0	5.5	93.4	1.1	7.9	90.6	1.6	1.0	95.0	4.0
v_{J1}^*	2.7	96.9	0.4	1.7	97.9	0.4	3.6	94.4	2.0	5.0	93.9	1.1	6.6	92.3	1.1	1.0	95.0	4.0
	srs epsem $m = 25$			srs epsem $m = 3$			srs epsem $m = 300$			srs epsem $m = 50$			srs epsem $m = 15$			srs epsem $m = 1{,}500$		
v_E	1.7	96.2	2.1	0.0	99.9	0.1	2.4	94.7	2.9	2.3	95.5	2.2	1.1	97.1	1.8	3.0	94.0	3.0
v_a	5.6	91.2	3.2	6.5	91.5	2.0	2.6	94.1	3.3	5.1	92.2	2.7	8.3	90.4	1.3	3.0	96.0	1.0
v_{wr}	5.8	91.2	3.0	9.6	87.2	3.2	3.1	93.3	3.6	3.4	95.1	1.5	9.3	89.7	1.1	3.0	95.0	2.0
v_{JL}	5.1	92.4	2.5	6.5	91.2	2.3	2.6	94.1	3.3	2.8	96.0	1.2	8.2	90.9	0.9	3.0	96.0	1.0
v_R	5.2	92.3	2.5	8.4	88.3	3.3	2.6	94.1	3.3	2.9	95.7	1.4	8.8	90.2	1.0	3.0	96.0	1.0
v_D	3.7	94.3	2.0	5.5	92.8	1.7	2.5	94.3	3.2	2.3	96.9	0.8	7.8	91.6	0.7	3.0	96.0	1.0
v_{J2}	1.9	97.3	0.8	2.6	96.7	0.7	2.3	94.9	2.8	2.0	97.9	0.1	6.9	92.6	0.5	3.0	96.0	1.0
v_{Jack}	2.2	96.8	1.0	4.7	94.0	1.3	2.3	94.9	2.8	2.1	97.8	0.1	7.3	92.1	0.6	3.0	96.0	1.0
v_{J1}	1.8	97.5	0.7	2.5	96.9	0.6	2.3	94.9	2.8	2.0	97.9	0.1	6.2	93.4	0.4	3.0	96.0	1.0
v_R^*	6.6	89.5	3.9	8.9	87.8	3.4	2.7	93.9	3.4	7.7	88.7	3.6	11.7	86.1	2.2	3.0	96.0	1.0
v_D^*	5.1	92.5	2.4	5.7	92.4	1.9	2.5	94.3	3.2	6.0	91.6	2.4	10.6	88.0	1.5	3.0	96.0	1.0
v_{J2}^*	3.4	94.9	1.7	2.8	96.5	0.7	2.3	94.9	2.8	4.6	93.7	1.7	9.2	89.7	1.1	3.0	96.0	1.0
v_{Jack}^*	3.5	94.8	1.7	4.9	93.7	1.4	2.3	94.9	2.8	4.7	93.3	2	9.9	89.0	1.2	3.0	96.0	1.0
v_{J1}^*	3.0	95.4	1.6	2.6	96.8	0.6	2.3	94.9	2.8	4.6	93.7	1.7	8.6	90.6	0.8	3.0	96.0	1.0

Table 3.5 (continued)
Coverage of 95% confidence intervals for population totals based on *t*-distributions and alternative variance estimators. See Table 3.1 for descriptions of the variance estimators

Variance est.	Third Grade			ACS			Simulation			Third Grade			ACS			Simulation		
	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper
	pps $m = 25$			*pps* $m = 3$			*pps* $m = 300$			*pps* $m = 50$			*pps* $m = 9$			*pps* $m = 1{,}500$		
v_E	1.7	95.9	2.4	0	100.0	0.0	2.9	94.2	2.9	2.3	95.3	2.4	0.7	98.0	1.3	2.0	95.0	3.0
v_a	6.2	90.0	3.8	4.7	94.3	1.0	2.9	93.9	3.2	3.1	94.1	2.8	5.1	94.4	0.5	2.0	92.0	6.0
v_{wr}	5.1	91.1	3.8	5.6	92.8	1.5	3.1	93.6	3.3	2.0	97.0	1.0	5.3	94.3	0.4	3.0	92.0	5.0
v_{JL}	4.9	92.0	3.1	4.9	93.5	1.5	2.9	94.0	3.1	1.9	96.9	1.2	4.9	94.7	0.3	2.0	92.0	6.0
v_R	5.3	91.5	3.2	7.2	90.5	2.3	2.9	93.9	3.2	2.0	96.8	1.2	5.6	94.1	0.4	2.0	92.0	6.0
v_D	3.8	94.1	2.1	4.4	94.4	1.1	2.7	94.7	2.6	1.7	97.4	0.9	4.8	94.9	0.3	2.0	92.0	6.0
v_{J2}	2.7	96.1	1.2	2.6	97.0	0.4	2.6	95.0	2.4	1.6	97.9	0.5	4.3	95.5	0.2	2.0	92.0	6.0
v_{Jack}	2.8	95.8	1.4	4.2	94.9	0.9	2.6	95.0	2.4	1.6	97.9	0.5	4.7	95.1	0.2	2.0	92.0	6.0
v_{J1}	2.2	96.7	1.1	2.1	97.5	0.4	2.6	95.0	2.4	1.5	98.0	0.5	3.9	96.0	0.1	2.0	92.0	6.0
v_R^*	7.4	87.8	4.8	7.6	90.0	2.4	2.9	93.9	3.2	5.0	90.6	4.4	8.9	89.8	1.3	2.0	92.0	6.0
v_D^*	5.3	91.6	3.1	4.7	94.0	1.3	2.7	94.5	2.8	4.1	92.2	3.7	8.1	90.9	1.0	2.0	92.0	6.0
v_{J2}^*	3.6	94.3	2.1	2.8	96.8	0.4	2.6	95.0	2.4	3.0	94.1	2.9	7.2	92.0	0.7	2.0	92.0	6.0
v_{Jack}^*	4.0	93.7	2.3	4.5	94.5	1.0	2.6	95.0	2.4	3.1	94.0	2.9	7.9	91.1	1.0	2.0	92.0	6.0
v_{J1}^*	3.5	94.6	1.9	2.2	97.4	0.4	2.6	95.0	2.4	2.9	94.4	2.7	6.8	92.6	0.6	2.0	92.0	6.0

One feature of v_D and v_D^* is that both the cluster-specific contributions, $v_{D,i}$ and $v_{D,i}^*$, as well as the overall variance estimates can be negative. In the simulations, the adjustment described after (2.11) was used to avoid negative contributions. Negative estimates were more common when the second stage sample sizes were small and the weights were quite variable. For example, for the ACS population, almost 28% of the simple random samples of 3 clusters and $m_i = 9$ resulted in at least one negative variance contribution for a cluster. More commonly, about 10% of the samples contained at least one negative variance estimate for a cluster. In the Third Grade population, 16% to 27% of the samples had at least one negative value of v_{Di}. In the simulated population with large sample sizes, v_{Di} was negative in less than 5% of the samples. With the ad hoc correction of setting $I_i - H_{ii}$ to I_i, v_D is one of the most attractive variance estimators because it tends to slightly overestimate the empirical variance, has some of the best confidence interval coverage, and has reasonable variability compared to other variance estimators.

4 Conclusion

Leverage adjustments to standard variance estimators have been shown to reduce bias and improve confidence interval coverage based on general regression estimators in single-stage samples. This paper extends those results to two-stage samples by presenting new adjustments based on hat matrices. Our theory provides the justification for the adjustments and illustrates that some of the proposed estimators are related to the delete-a-cluster jackknife that is a common procedure in survey estimation.

To test the theory, we conducted a series of simulation studies on three populations designed to assess performance in a variety of situations. In a school population a large sampling fraction of first-stage units was used. In a second population, based on American Community Survey data, the effects of small sample

sizes were tested. In a third simulated population, we examined large sample performance. Both simple random sampling and probability proportional to size sampling of clusters were used.

The relationships of the variance estimators were similar across all sample designs. The with-replacement variance estimator, υ_{wr}, which is the default choice in survey software packages, the jackknife linearization estimator, υ_{JL}, and the design-based variance estimator, υ_g, that assumes Poisson sampling at each stage as a computational convenience, are often negatively biased leading to confidence intervals that cover at less than the desired rate. Some of the jackknife-related estimators – υ_{Jack}, υ_{J1}, and υ_{J2} – which explicitly or implicitly include hat-matrix adjustments, are prone to producing large, outlying values when the first-stage sample is small. This is especially true when the first-stage is selected by *srs* but is less so in *pps* sampling when an efficient measure of size is used.

The variance estimators proposed here, particularly υ_D, provide alternatives to estimating the variance of GREG estimators in complex samples. At the expense of somewhat inflating the variability of the variance estimator, the hat-matrix adjusted sandwich estimators, denoted here by v_D, v_{J1}, and v_{J2}, give confidence interval coverage that is closer to the nominal value in small to moderate samples. Depending on the sample design and population characteristics, hat-matrix adjusted estimators can produce less biased variance estimates and better inferences when compared to the standard methods.

Acknowledgements

The authors thank the associate editor and two referees whose comments substantially improved the presentation.

Appendix

Theoretical results

A.1 Assumptions

The assumptions used to obtain asymptotic results are listed below. The number of population and sample clusters approach infinity; however, the number of population clusters increases at a faster rate than the number of sample clusters. Certain population quantities are assumed to be bounded.

A.1.1 $m/M \to 0$ *as* $m \to \infty$ *and* $M \to \infty$.

A.1.2 *All* N_i *and* n_i *are bounded.*

A.1.3 $\pi_{ik} = O(m/M)$ *for all* ik.

A.1.4 *All elements of* $\mathbf{X}$, $\mathbf{\Psi}$, *and* $\mathbf{Q}$ *are bounded.*

A.1.5 *The sample design is such that* $\frac{\sqrt{m}}{M}\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_{Ux}\right) \xrightarrow{d} N(0, \mathbf{V})$, *where* $\mathbf{V}$ *is a* $p \times p$ *positive definite matrix, i.e.,* $\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_{Ux}\right) = O_p\left(M/\sqrt{m}\right)$.

Since $\mathbf{\Pi} = O\left(\frac{m}{M}\right)$ elementwise and $\mathbf{A} = \mathbf{X}_s^\top \mathbf{Q}^{-1} \mathbf{\Pi}^{-1} \mathbf{X}_s$ can be written as the sum of n terms and n_i is bounded while $m \to \infty$, $\mathbf{A} = O(M)$. By definition $\mathbf{g}_i^\top = \mathbf{1}_{n_i} + \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right)^\top \mathbf{A}^{-1} \mathbf{X}_i^\top \mathbf{Q}_i$. The second term in g_i is $O_p\left(m^{-1/2}\right)$; consequently $\mathbf{g}_i$ converges to a vector of 1's. Using $\mathbf{A} = O(M)$ along with assumptions A.1.3 and A.1.4, $\mathbf{H}_{ij}$ is $O\left(m^{-1}\right)$ elementwise.

A.2 Model variance of GREG

Let $\mathbf{y}_{si}$ be the vector of all sample elements in cluster i and $\mathbf{y}_i$ be the vector of all elements in cluster i. The variance of the GREG, with respect to the working model (2.1) is:

$$\begin{aligned}
\mathrm{var}_\xi\left(\hat{t}_y^{gr} - t_y\right) &= \mathrm{var}_\xi\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{y}_{si} - \sum_{i\in U} \mathbf{1}_{N_i}^\top \mathbf{y}_i\right) \\
&= \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{\Psi}_{si} \mathbf{\Pi}_i^{-1} \mathbf{g}_i - 2\mathrm{cov}_\xi\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{y}_{si}, \sum_{i\in U} \mathbf{1}_{N_i}^\top \mathbf{y}_i\right) + \mathbf{1}_N^\top \mathbf{\Psi} \mathbf{1}_N.
\end{aligned}$$

Since $\sum_{i\in U} \mathbf{1}_i^\top \mathbf{y}_i = \sum_{i\in s} \mathbf{1}_i^\top \mathbf{y}_i + \sum_{i\in(U-s)} \mathbf{1}_i^\top \mathbf{y}_i$ and elements in different clusters are uncorrelated, we have,

$$\begin{aligned}
\mathrm{var}_\xi\left(\hat{t}_y^{gr} - t_y\right) &= \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{\Psi}_{si} \mathbf{\Pi}_i^{-1} \mathbf{g}_i - 2\sum_{i\in s}\left[\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathrm{cov}_\xi\left(\mathbf{y}_{si}, \mathbf{y}_i\right) \mathbf{1}_{N_i}\right] + \mathbf{1}_N^\top \mathbf{\Psi} \mathbf{1}_N \\
&= L_1 - 2L_2 + L_3.
\end{aligned}$$

Since $\mathbf{A}^{-1} = O\left(M^{-1}\right)$ and $\mathbf{g}_i$ and $\mathbf{\Psi}_{si}$ are bounded, we have $L_1 = O\left(M^2/m\right)$. Because $\mathbf{\Psi}_{si}$ is bounded, $\mathrm{cov}_\xi\left(\mathbf{y}_{si}, \mathbf{y}_i\right) = O(1)$ and $L_2 = O(M)$. L_3 is the sum of N terms. Since the N_i are bounded, $L_3 = O(M)$. Thus, L_1 is the dominant term of the prediction variance.

A.3 Proof that $\mathrm{var}_\xi\left(\mathbf{e}_i\right) \approx \mathbf{\Psi}_{si}$

In this section in order to simplify the notation, we omit the subscript s on $\mathbf{y}_{si}$, $\hat{\mathbf{y}}_{si}$, and $\mathbf{\Psi}_{si}$. The residual can be written in terms of a hat matrix as follows.

$$\begin{aligned}
\mathbf{e}_i &= \mathbf{y}_i - \hat{\mathbf{y}}_i \\
&= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\mathbf{y}_i - \sum_{j\neq i;\, i,\, j\in s} \mathbf{H}_{ij} \mathbf{y}_j
\end{aligned}$$

where $\mathbf{I}_{n_i}$ is the $n_i \times n_i$ identity matrix. The model variance of $\mathbf{e}_i$ is then

$$\begin{aligned}
\mathrm{var}_\xi\left(\mathbf{e}_i\right) &= \mathrm{var}_\xi\left[\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\mathbf{y}_i - \sum_{j\neq i} \mathbf{H}_{ij} \mathbf{y}_j\right] \\
&= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right) \mathrm{var}_\xi\left(\mathbf{y}_i\right)\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i} \mathbf{H}_{ij} \mathrm{var}_\xi\left(\mathbf{y}_j\right) \mathbf{H}_{ij}^\top \\
&= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right) \mathbf{\Psi}_i \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i} \mathbf{H}_{ij} \mathbf{\Psi}_j \mathbf{H}_{ij}^\top. \qquad \text{(A.1)}
\end{aligned}$$

As noted above, $\mathbf{H}_{ii} = O\left(m^{-1}\right)$. Thus, $\mathrm{var}_\xi\left(\mathbf{e}_i\right) = \mathbf{\Psi}_i + O\left(m^{-1}\right)$.

To justify υ_D, note that the second term of (A.1) can be written as

$$\sum_{j\neq i} \mathbf{H}_{ij} \mathbf{\Psi}_j \mathbf{H}_{ij}^\top = \sum_{j\in s} \mathbf{H}_{ij} \mathbf{\Psi}_j \mathbf{H}_{ij}^\top - \mathbf{H}_{ii} \mathbf{\Psi}_i \mathbf{H}_{ii}^\top.$$

The sum over the full cluster sample is

$$\sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = \mathbf{X}_i\mathbf{A}^{-1}\left(\sum_{j\in s}\mathbf{X}_j^\top\mathbf{Q}_j\boldsymbol{\Pi}_j^{-1}\boldsymbol{\Psi}_j\boldsymbol{\Pi}_j^{-1}\mathbf{Q}_j\mathbf{X}_j\right)\mathbf{A}^{-1}\mathbf{X}_i^\top.$$

In the special case of $\mathbf{Q}_j = \boldsymbol{\Psi}_j^{-1}$ and $\boldsymbol{\Pi}_i = c\mathbf{I}_{n_i}$ for some constant $c \in (0, 1)$ (i.e., the sample is self-weighting), we have

$$\sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = c^{-2}\mathbf{X}_i\mathbf{A}^{-1}\left(\sum_{j\in s}\mathbf{X}_j^\top\boldsymbol{\Psi}_j^{-1}\mathbf{X}_j\right)\mathbf{A}^{-1}\mathbf{X}_i^\top,$$

along with $\mathbf{H}_{ii} = c\mathbf{X}_i\mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Psi}_i^{-1}$ and $\mathbf{A} = c^{-1}\mathbf{X}\boldsymbol{\Psi}^{-1}\mathbf{X}$. Using these simplifications, $\sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = \mathbf{H}_{ii}\boldsymbol{\Psi}_i$. Substituting this result in (A.1) and simplifying gives

$$\begin{aligned}\text{var}_\xi(\mathbf{e}_i) &= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\boldsymbol{\Psi}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top \\ &= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\boldsymbol{\Psi}_i. \qquad \text{(A.2)}\end{aligned}$$

This is the basis for the adjustment of υ_R to obtain υ_D.

A.4 Proof that $\hat{\mathbf{B}}_{(i)} = \hat{\mathbf{B}} - \mathbf{R}_i$ for cluster samples

In this section, we omit the subscript s on $\mathbf{X}_s$, $\mathbf{y}_s$, $\mathbf{X}_{si}$, $\mathbf{y}_{si}$, $\mathbf{X}_{s(i)}$, and $\mathbf{y}_{s(i)}$ to simplify the notation. The subscript (i) denotes removal of the i^{th} cluster from the full sample matrix or vector. For example, $\hat{\mathbf{B}}_{(i)}$ is an estimate of $\mathbf{B}$ based on all sample clusters except cluster i and is

$$\hat{\mathbf{B}}_{(i)} = \left(\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)}$$

where $\mathbf{W}_{(i)} = \mathbf{Q}_{(i)}\boldsymbol{\Pi}_{(i)}^{-1}$. Using Lemma 9.5.1 in Valliant et al. (2000), we have

$$\hat{\mathbf{B}}_{(i)} = \left(\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{X}_i\mathbf{A}^{-1}\right)\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)}.$$

Since $\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)} = \mathbf{X}^\top\mathbf{W}\mathbf{y} - \mathbf{X}_i^\top\mathbf{W}_i\mathbf{y}_i$ and $\hat{\mathbf{B}} = \mathbf{A}^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{y}$, we have

$$\begin{aligned}\hat{\mathbf{B}}_{(i)} &= \mathbf{A}^{-1}\left(\mathbf{X}^\top\mathbf{W}\mathbf{y} - \mathbf{X}_i^\top\mathbf{W}_i\mathbf{y}_i\right) \\ &\quad + \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{X}_i\mathbf{A}^{-1}\left(\mathbf{X}^\top\mathbf{W}\mathbf{y} - \mathbf{X}_i\mathbf{W}_i\mathbf{y}_i\right) \\ &= \hat{\mathbf{B}} - \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\mathbf{y}_i + \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\hat{\mathbf{y}}_i \\ &\quad - \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{H}_{ii}\mathbf{y}_i \\ &= \hat{\mathbf{B}} - \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i.\end{aligned}$$

That is, $\hat{\mathbf{B}}_{(i)} = \hat{\mathbf{B}} - \mathbf{R}_i$.

A.5 Jackknife variance estimator of clustered GREG in terms of leverages

We now simplify the delete-a-cluster Jackknife variance estimator of the clustered GREG. As in Sections A.3 and A.4, we omit the subscript s on various terms to simplify the notation. The estimated total after removing the i^{th} cluster is defined as

$$\begin{aligned}
\hat{t}_{y(i)}^{gr} &= \frac{m}{m-1}\hat{t}_{y(i)}^{\pi} + \left[\mathbf{t}_{Ux} - \frac{m}{m-1}\hat{\mathbf{t}}_{x(i)}^{\pi}\right]\hat{\mathbf{B}}_{(i)} \\
&= \frac{m\mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{y}}{m-1} - \frac{m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{y}_i}{m-1} + \left[\mathbf{1}_N^\top \mathbf{X}_U - \frac{m\mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{X}}{m-1} + \frac{m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i}{m-1}\right]\left(\hat{\mathbf{B}} - \mathbf{R}_i\right) \\
&= \frac{m\mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{y}}{m-1} - \frac{m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{y}_i}{m-1} \\
&\quad + \frac{m}{m-1}\left(\mathbf{1}_N^\top \mathbf{X}_U - \mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{X}\right)\left(\hat{\mathbf{B}} - \mathbf{R}_i\right) - \frac{1}{m-1}\left(\mathbf{1}_N^\top \mathbf{X}_U - m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i\right)\left(\hat{\mathbf{B}} - \mathbf{R}_i\right) \\
&= \frac{m}{m-1}\hat{t}_y^{gr} - \frac{m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{y}_i}{m-1} - \frac{m}{m-1}\left(\mathbf{1}_N^\top \mathbf{X}_U - \mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{X}\right)\mathbf{R}_i - \frac{1}{m-1}K_i.
\end{aligned}$$

Adding and subtracting $\frac{m}{m-1}\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i$ and doing a substantial amount of simplification leads to

$$\hat{t}_{y(i)}^{gr} = \frac{m}{m-1}\hat{t}_y^{gr} - \frac{m}{m-1}\mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i + \frac{m}{m-1}G_i - \frac{1}{m-1}K_i.$$

Taking the difference between the delete-one estimates and the average of those estimates gives

$$\begin{aligned}
\hat{t}_{y(i)}^{gr} - \hat{t}_{y(\cdot)}^{gr} &= -\frac{m}{m-1}\left(D_i - \bar{D}\right) + \frac{m}{m-1}\left(G_i - \bar{G}\right) - \frac{1}{m-1}\left(K_i - \bar{K}\right) \\
&= -\frac{m}{m-1}\left(D_i - \bar{D}\right) + \frac{m}{m-1}\left[\left(G_i - \bar{G}\right) - \frac{1}{m}\left(K_i - \bar{K}\right)\right].
\end{aligned}$$

Letting $F_i = \left(G_i - \bar{G}\right) - m^{-1}\left(K_i - \bar{K}\right)$ leads to the formula for υ_{Jack} in equation (2.12). Next, since $\mathbf{H}_{ii} = O\left(m^{-1}\right)$ and $\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\mathbf{B}}$,

$$\begin{aligned}
F_i &= \left(G_i - \bar{G}\right) - \frac{1}{m}\left(K_i - \bar{K}\right) \\
&\approx \left[-\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\hat{\mathbf{y}}_i + \frac{1}{m}\sum_{i\in s}\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\hat{\mathbf{y}}_i\right] - \frac{1}{m}\left[-m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i\hat{\mathbf{B}} + \sum_{i\in s}\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i\hat{\mathbf{B}}\right] \\
&= \mathbf{0}.
\end{aligned}$$

Thus, $F_i = o(1)$, and υ_{Jack} in (2.6) and (2.12) is asymptotically equivalent to υ_{J1} in (2.13).

Finally, to justify υ_{J2} in (2.14), we write υ_{J1} in the computational form

$$\upsilon_{J1} = \frac{m}{m-1}\left[\sum_{i\in s}\left(\mathbf{g}_i^\top U_i\mathbf{e}_i\right)^2 - \frac{1}{m}\left(\sum_{i\in s}\mathbf{g}_i^\top U_i\mathbf{e}_i\right)^2\right] \tag{A.3}$$

where $U_i = \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}$. Note that the model variance of D_i is

$$\begin{aligned}
\text{var}_\xi\left(D_i\right) &= \text{var}_\xi\left(\mathbf{g}_i^\top U_i\mathbf{e}_i\right) \\
&= \mathbf{g}_i^\top U_i^\top \text{var}_\xi\left(\mathbf{e}_i\right)U_i\mathbf{g}_i.
\end{aligned}$$

Because $U_i = O\left(M/m\right)$ and the sum in $\sum_{i\in s}\text{var}_\xi\left(D_i\right)$ contains $n = m\bar{n}$ terms, the variance of $\sum_{i\in s}\mathbf{g}_i^\top U_i\mathbf{e}_i$ is $O\left(M^2/m\right)$. Next, scaling υ_{J1} to be appropriate for a mean, the first term in the brackets

in (A.3) is $N^{-2}\sum_{i\in s} D_i^2 = O(m^{-1})$. Since the second term in brackets has model expectation 0 and variance that is $O(m^{-1})$, it converges in probability to 0, and v_{J2} is asymptotically equivalent to v_{J1}.

A.6 Asymptotic equivalence of variance estimators

In this appendix we sketch arguments for why several variance estimators are asymptotically equivalent. Using design-based arguments, Yung and Rao (1996, Appendix) showed that the jackknife linearization estimator, v_{JL}, for the GREG is asymptotically equivalent to the design-consistent estimator, v_{Jack}, in stratified multistage designs with a large number of strata and a bounded number of sample clusters selected from each stratum. Using regularity conditions in Rao and Shao (1985), that result can be extended to cover designs in which either (i) the number of strata is large and the number of clusters per stratum is bounded or (ii) the number of strata is limited and the number of sample clusters per stratum is large, as is the case in this article.

The jackknife linearization estimator in Section 2 can be expanded as

$$N^{-2}v_{JL} = N^{-2}\sum_{i\in s}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{e}_i\mathbf{e}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{g}_i - N^{-2}m\left(m^{-1}\sum_{i\in s}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{e}_i\right)^2. \quad \text{(A.4)}$$

The first term in (A.4) equals v_R. Because, under some reasonable assumptions, $\mathbf{g}_i$ and $\mathbf{e}_i$ are bounded, and $\mathbf{\Pi}_i^{-1} = O(M/m)$ by assumptions A.1.2 and A.1.3, the first term in (A.4) is $O(1/m)$. The second term is also $O(1/m)$, but the model expectation of $\bar{\mathbf{e}}_2 = m^{-1}\sum_{i\in s}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{e}_i$ is zero as long as (2.1) holds. Since $\bar{\mathbf{e}}_2$ is a mean, its model-variance will approach 0 as $m\to\infty$. Thus, the second term in (A.4) will converge in probability to 0 and $v_{JL}\approx v_R$.

In Section A.5 it was shown that v_{Jack} and v_{J1} are asymptotically equivalent. Under A.1.1-A.1.4, $\mathbf{H}_{ii} = O(m^{-1})$. Consequently, v_{J2} and v_D are approximately the same as v_R as $m\to\infty$. Thus, $v_{\text{Jack}}\approx v_{JL}$ by extension of Yung and Rao (1996), both of which are design-consistent. Further, v_{JL} is asymptotically equivalent to v_{J1}, v_{J2}, v_D, and v_R. As a result, the alternative variance estimators considered here all have both model-based and design-based justifications.

References

Kott, P.S. (1988). Model-based finite population correction for the Horvitz-Thompson estimator. *Biometrika*, 75(4), 797-799.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9(5), 1010-1019.

Li, J., and Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, 35, 1, 15-24. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10881-eng.pdf.

Long, J.S., and Ervin, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3), 217-224.

MacKinnon, J.G., and White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3), 305-325.

Rao, J.N.K., and Shao, J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80(391), 620-630.

Royall, R.M., and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73(362), 351-358.

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

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

Valliant, R. (2002). Variance estimation for the general regression estimator. *Survey Methodology*, 28, 1, 103-114. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002001/article/6424-eng.pdf.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley Series in Probability and Statistics: Survey Methodology Section. New York: John Wiley & Sons, Inc.

Yung, W., and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 1, 23-31. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996001/article/14388-eng.pdf.