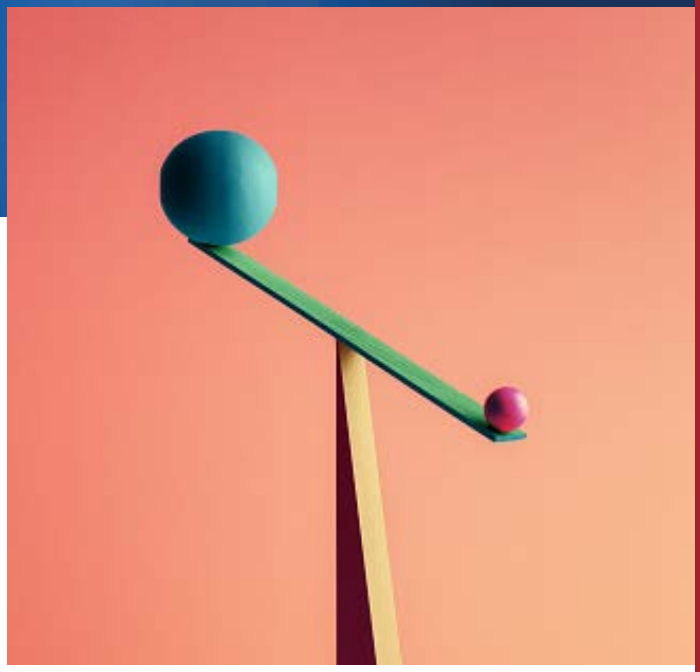
RTI Press

Methods Report

ISSN 2378-7813

March 2022



The Role of Weights in Regression Modeling and Imputation

Phillip S. Kott

RTI INTERNATIONAL

RTI Press publication MR-0047-2203

RTI International is an independent, nonprofit research organization dedicated to improving the human condition. The RTI Press mission is to disseminate information about RTI research, analytic tools, and technical expertise to a national and international audience. RTI Press publications are peer-reviewed by at least two independent substantive experts and one or more Press editors.

Suggested Citation

Kott, P. S. (2022). *The Role of Weights in Regression Modeling and Imputation*. RTI Press Publication No. MR-0047-2203. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2022.mr.0047.2203

This publication is part of the RTI Press Methods Report series.

RTI International
3040 East Cornwallis Road
PO Box 12194
Research Triangle Park, NC
27709-2194 USA

Tel: +1.919.541.6000
E-mail: rtipress@rti.org
Website: www.rti.org

©2022 RTI International. RTI International is a trade name of Research Triangle Institute. RTI and the RTI logo are U.S. registered trademarks of Research Triangle Institute.

This work is distributed under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 license (CC BY-NC-ND), a copy of which is available at https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode

Contents

About the Author

Phillip S. Kott, PhD, is a senior research statistician in RTI International's Center of Excellence for Complex Data Analysis.

RTI Press Associate Editor

Valerie Williams

Acknowledgments

I would like to thank Valerie Williams, the RTI Press associate editor who managed peer-review, and the reviewers whose thoughtful comments made this a stronger paper.

Abstract

When fitting observations from a complex survey, the standard regression model assumes that the expected value of the difference between the dependent variable and its model-based prediction is zero, regardless of the values of the explanatory variables. A rarely failing extended regression model assumes only that the model error is uncorrelated with the model's explanatory variables. When the standard model holds, it is possible to create alternative analysis weights that retain the consistency of the model-parameter estimates while increasing their efficiency by scaling the inverse-probability weights by an appropriately chosen function of the explanatory variables.

When a regression model is used to impute for missing item values in a complex survey and when item missingness is a function of the explanatory variables of the regression model and not the item value itself, near unbiasedness of an estimated item mean requires that either the standard regression model for the item in the population holds or the analysis weights incorporate a correctly specified and consistently estimated probability of item response. By estimating the parameters of the probability of item response with a calibration equation, one can sometimes account for item missingness that is (partially) a function of the item value itself.

Introduction

When fitting a regression model with complex survey data, one frequently treats the finite population as a realization of independent trials from a conceptual population and tries to use the complex sample to estimate with probability-sampling principles either a maximum likelihood (ML) estimator computed from the finite population or the limit of the putative estimator as the population grows arbitrarily large (see Fuller, 1975 for linear regression and Binder, 1983 more generally).

We do not take that "design-based" approach here. Instead, we adopt a model-based framework from Kott (2007, 2018). This framework is *sensitive* to the complex sampling design and to the possibility that many of the usual model assumptions may not hold in the population. Under this design-sensitive framework, some methods developed in the conventional design-based framework are retained, such as fitting weighted estimating equations and sandwich variance/mean-squared-error estimation, but their interpretations change.

I begin this report by laying out the design-sensitive approach to regression modeling with complex survey data, which involves distinguishing between the robust *standard* model and the more general *extended* model. Estimating model parameters under the extended model requires the use of inverse-probability weights (broadly defined here to include calibration adjustments), although those weights may not be needed under the standard model. Even when such weights are helpful under the standard model, it may be possible to modify them to increase the efficiency of parameter estimates. Then, I offer a description of variance estimation followed by a reference to some useful tests for determining whether using inverse-probability weights is necessary and whether the standard model holds in the population.

My focus then changes to using a standard regression model to impute for missing item values in an estimated total (or mean) by first assuming an item-response model where item nonresponse is missing at random. This methodology is extended to situations where item nonresponse is not missing at random, providing a nearly unbiased estimate for an item mean in some sense when the standard model fails and a more efficient estimate when it does not.

I conclude the paper with a review of the ideas developed here and add speculations about imputation.

The Design-Sensitive Approach to Regression Modeling

Following Kott (2018), the *standard regression model* assumes that given any element (member) k of a population U,

$$y_k = f(\mathbf{z}_k{}^T\boldsymbol{\beta}) + \varepsilon_k, \quad (1)$$

where

$$\mathrm{E}(\varepsilon_k \,|\mathbf{z}_k) = 0 \text{ for all } \mathbf{z}_k,\ k \in U \quad (2)$$

In Equation (1), y_k is the dependent random variable being modeled whereas $\mathbf{z}_k$ is a vector of P explanatory variables (covariates), one of which is 1 or the equivalent (some linear combination of the components of $\mathbf{z}_k$ is 1 for all $k \in U$), and $f(.)$ is a specified monotonic function. In particular, $f(\mathbf{z}_k{}^T\boldsymbol{\beta}) = \mathbf{z}_k{}^T\boldsymbol{\beta}$ for a linear regression model whereas $f(\mathbf{z}_k{}^T\boldsymbol{\beta}) = \exp(\mathbf{z}_k{}^T\boldsymbol{\beta})/[1 + \exp(\mathbf{z}_k{}^T\boldsymbol{\beta})]$ for a logistic regression model, where $\boldsymbol{\beta}$ is an unknown vector of parameters that can be estimated using a sample drawn from U. Some of the components of $\mathbf{z}_k$ can be random variables.

Poisson regression, where $f(\mathbf{z}_k{}^T\boldsymbol{\beta}) = \exp(\mathbf{z}_k{}^T\boldsymbol{\beta})$, is often assumed when the dependent variable is restricted to positive values. This restriction can be extended to positive integers. In practice, Poisson regression often multiplies $\exp(\mathbf{z}_k{}^T\boldsymbol{\beta})$ by a known offset variable o_k. For our purposes, this offset variable can be thought of as being incorporated into a revised dependent variable: $y_k{}^\star = y_k/o_k$.

Few additional assumptions about the distribution and variance structure of the ε_k are needed in the above broadly specified version of the model until the issue of estimating the variance of an estimator for β arises. That is a subject I take up shortly.

A restriction imposed by the standard model in Equation (1) is that the expected value of the error

 https://doi.org/10.3768/rtipress.2022.mr.0047.2203

term ε_k is 0 no matter the value of $\mathbf{z}_k$. This assumption can fail. A generalization of the standard model is the extended model under which $E(\varepsilon_k \mid \mathbf{z}_k) = 0$ in Equation (2) is replaced by

$$E(\mathbf{z}_k\varepsilon_k) = \mathbf{0}. \quad (3)$$

In other words, ε_k need only have mean 0 unconditionally (i.e., $E(\varepsilon_k) = 0$) rather than when conditioned on $\mathbf{z}_k$ for any $\mathbf{z}_k$. Unconditional unbiasedness obtains because 1 is either a component of $\mathbf{z}_k$ or a linear combination of the components of $\mathbf{z}_k$. Equation (3) simply requires ε_k to be uncorrelated with any random components of $\mathbf{z}_k$. Unlike the standard model, the more general extended model rarely fails, as long as the first three central moments of the components of $\mathbf{z}_k$ are finite (formally, this means that as the population size M grows arbitrarily large, the limit of each component of $\bar{\mathbf{z}}$ and of $M^{-1}\sum_U(\mathbf{z}_k - \bar{\mathbf{z}})^r$, where $\bar{\mathbf{z}} = M^{-1}\sum_U \mathbf{z}_k$ and $r = 2$ or 3, is finite).

Observe that the standard version of the simple linear model through the origin, $y_k = \beta z_k + \varepsilon_k$, is not exactly of the form specified by Equation (1) because it is missing an intercept. It similarly assumes $E(\varepsilon_k \mid z_k) = 0$. The extended version of this model assumes only $E(\varepsilon_k) = 0$.

The Group-Mean and Ratio Models

Suppose the population U can be divided into G mutually exclusive and exhaustive groups. Let $\boldsymbol{\delta}_k = (\delta_{k1}, \delta_{k2} \ldots, \delta_{kG})^T$, where $\delta_{kg} = 1$ when element k is in the g^{th} group and 0 otherwise. Let us now investigate the linear regression model:

$$y_k = (q_k\boldsymbol{\delta}_k{}^T)\,\boldsymbol{\beta} + \varepsilon_k, \quad (4)$$

where q_k is a scalar, and $E(\varepsilon_k|\boldsymbol{\delta}_k) = 0$. When $q_k \equiv 1$ (or, equivalently, any other constant), Equation (4) is called the *group-mean model*, because the mean of every element in group g is the same: β_g. When the q_k vary within groups, Equation (4) is called the group-ratio model. This is a useful model in business surveys where q_k is often a measure of size known for all elements in the population.

When $G = 1$, the group-mean model devolves into the population-mean model and the group-ratio model devolves into the population-ratio model. When $G > 1$ and $q_k \equiv 1$ in Equation (4), the value β_g is the mean of g^{th} group, also called the domain mean of group g.

Unlike the group-mean model, the group-ratio model does not fit our formal definition of a regression model in Equation (1) because $\mathbf{z}_k = \boldsymbol{\delta}_k q_k$ does not contain 1 among its components or the equivalent unless q_k is a constant for $k \in U$. Equation (2) is effectively replaced by $E(\varepsilon_k \mid \boldsymbol{\delta}_k) = 0$ for all realized $\boldsymbol{\delta}_k$, $k \in U$.

As long as the y_k are bounded, neither the group-mean model nor the group-ratio model fails. Although the assumption that the y_k are bounded may seems reasonable, a referee correctly pointed out that this assumption means that the y_k cannot be normally distributed. Rather than formulating an asymptotic framework where, as B grows arbitrarily large, the probability that $|y_k|$ is greater than B tends toward 0 at an appropriate speed, let us concede the referee's point and argue that in a finite world, the assumption that each y_k is normally distributed is a never-realized idealization.

The Weighted Estimating Equation

For now, we will mostly restrict our attention here to probability samples. This means that every $k \in U$ has a positive probability π_k of being selected into the sample. Formally, $\pi_k \geq B_\pi > 0$ for some B_π.

Although populations from which probability samples are drawn are almost always finite, the samples themselves are often large. That is why it is reasonable to use asymptotics (arbitrarily large sample properties) when analyzing probability-sample data. Moreover, when modeling a finite population, we are less interested in the population itself than in a mechanism that can be hypothesized to have generated that population and could continue to generate elements *ad infinitum*.

Consequently, we assume there is an infinite sequence of nested populations growing arbitrarily large and that a sample can be drawn from each using the same probability-sampling mechanism. The samples in the sequence of samples, although not necessarily nested within each other, also grow arbitrarily large. As a result, it is possible to take the probability limit of a statistic based on a sample as the expected number

of sampled elements grows arbitrarily large (as we advance from one population in the sequence of populations to the next *ad infinitum*).

Suppose t_y is an estimator for the population total T_y. A sufficient condition for the probability limit of t_y, which we denote $p\lim(t_y)$, to be T_y as the population and sample sizes grow arbitrarily large is for the limit of the relative mean-squared error of t to converge to 0. When that happens, t_y is a consistent estimator for T_y.

Letting M denote the number of elements in population U, it is not difficult to see that

$$p\lim\left\{M^{-1}\sum_{k\in U}\mathbf{z}_k\left[y_k - f(\mathbf{x}_k{}^T\boldsymbol{\beta})\right]\right\} = p\lim\left\{M^{-1}\sum_{k\in U}\mathbf{z}_k\varepsilon_k\right\} = 0 \quad (5)$$

under the extended model (where $\mathrm{E}(\mathbf{z}_k\varepsilon_k) = \mathbf{0}$) with mild assumptions about the values of the components of $\mathbf{z}_k$ (e.g., they are bounded in number, and each have finite moments) and the variance structure of the ε_k (which we will discuss in some detail shortly).

Given a probability sample S with *analysis weights* $\{w_k\}$, each (nearly) equal to the $1/\pi_k$,

$$p\lim\left\{M^{-1}\sum_{k\in S}w_k\mathbf{z}_k\left[y_k - f(\mathbf{x}_k{}^T\boldsymbol{\beta})\right]\right\} = 0 \quad (6)$$

under mild additional conditions on the sampling design and population such that

$$p\lim\Psi_q = 0, \quad (7)$$

where

$$\Psi_q = M^{-1}\left(\sum_S w_k q_k - \sum_U q_k\right),$$

and q_k can equal 1, y_k, a component of $\mathbf{z}_k$, or a product of the previous variables.

Sufficient additional assumptions include that each of the q_k have finite moments and that the sample size grows arbitrarily large along with the population. I will make more assumptions about the sample design shortly.

Two sample-based values are said to be nearly equal when their ratio tends to 1 (in probability) as the sample size grows arbitrarily large. Similarly, an estimator is nearly unbiased when its relative bias tends to 0 as the sample size grows.

The analysis weights w_k may not be exactly equal to the $1/\pi_k$. Sometimes, analysis weights are calibrated to increase the statistical efficiency of the resulting estimators (as in Deville and Särndal, 1992) or to account for unit nonresponse or frame under- or over-coverage (e.g., Kott, 2006). Except in the forthcoming discussion on variance estimation via linearization, we treat the w_k as nearly equal to the inverse of the probability that element k is jointly in the frame, selected for the sample, and a sample respondent. We ignore the possibility of duplications in the frame. We treat S as the respondent sample and set $w_k = 0$ when $k \notin S$. For now, we assume there is no item nonresponse.

The w_k are inserted into Equation (6) in case $\mathrm{E}(\varepsilon_k \mid w_k) \neq 0$, a situation in which the analysis weights are said to be nonignorable in expectation (with respect to the model—a phrase that usually goes without saying). Full ignorability of the analysis weights or, equivalently, of the selection probabilities in the sense of Little and Rubin (2002), obtains when the conditional ε_k are independent of the w_k. Observe that if the original random sample is selected with probability proportion to some component of $\mathbf{z}_k$, while the variance of ε_k is a function of that same component, then ε_k is clearly not independent of w_k, and the weights are not ignorable, but they could still be ignorable in expectation (i.e., $\mathrm{E}(\varepsilon_k \mid w_k) = 0$ for every realized w_k, $k \in U$).

Whether the standard or extended model is assumed to hold in the population, solving for $\mathbf{b}$ in the weighted estimating equation (Godambe & Thompson, 1974)

$$\sum_{k\in S}w_k\mathbf{z}_k\left[y_k - f(\mathbf{z}_k{}^T\mathbf{b})\right] = 0 \quad (8)$$

provides a consistent estimator for $\boldsymbol{\beta}$ under mild conditions because

$$\mathbf{b} - \boldsymbol{\beta} = \left[M^{-1}\sum_{k\in S}w_k f'(\theta_k)\,\mathbf{z}_k\mathbf{z}_k{}^T\right]^{-1}M^{-1}\sum_{k\in S}w_k\mathbf{z}_k\varepsilon_k, \quad (9)$$

for some θ_k between $\mathbf{z}_k{}^T\mathrm{b}$ and $\mathbf{z}_k{}^T\beta$. This is a consequence of the mean-value theorem (see Kott, 2015 for an elaboration). An additional mild condition we assume is that

$$\mathbf{A}_\theta = M^{-1}\sum_{k\in S}w_k f'(\theta_k)\,\mathbf{z}_k\mathbf{z}_k{}^T \quad (10)$$

$$\mathbf{A} = M^{-1} \sum_{k \in S} w_k f'(\mathbf{z}_k{}^T \mathbf{b})\, \mathbf{z}_k \mathbf{z}_k{}^T \qquad (11)$$

and their probability limit, $\mathbf{A}^*$, has finite components and is positive definite. When $M^{-1}{}_S\, w_k \mathbf{z}_k \varepsilon_k$ converges to 0 in probability as the sample size grows arbitrarily large, $\mathbf{b}$ is a consistent estimator for $\boldsymbol{\beta}$.

It is not hard to show that $\sum_U \mathbf{z}_k[y_k - f(\mathbf{z}_k{}^T\mathbf{b})] = \mathbf{0}$ is the ML estimating equation for the population under the independent and identically distributed linear regression model and under logistic regression with independently sampled population elements. Nevertheless, the solution to Equation (8) is not ML when the weights vary or the ε_k within primary sampling units are correlated. Instead, the $\mathbf{b}$ solving Equation (8) is referred to as a *pseudo-ML* estimator for $\boldsymbol{\beta}$ (Skinner, 1989).

Pseudo-ML and Pfeffermann-Sverchkov Weight Adjustment

The pseudo-ML estimating equation in Binder (1983) is

$$\sum_{k \in S} w_k \left(\frac{f'(\mathbf{z}_k{}^T \mathbf{b})}{v_k} \right) \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T \mathbf{b}) \right] = \mathbf{0}. \qquad (12)$$

It derives from being the probability-sampling analog of the ML estimating equation when $v_k = \mathrm{E}(\varepsilon_k{}^2 | \mathbf{z}_k)$ is known (up to a scaling constant), and $\mathrm{E}(\varepsilon_k\, \varepsilon_j\, | \mathbf{z}_k, \mathbf{z}_j) = 0$ for $k \neq j$: $\sum_U \left(f'(\mathbf{z}_k{}^T \mathbf{b})/v_k \right) \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T \mathbf{b}) \right] = \mathbf{0}$. For ordinary least squares linear regression: $f'(\mathbf{z}_k{}^T\boldsymbol{\beta}) = 1$; for ordinary logistic regression: $f'(\mathbf{z}_k{}^T\boldsymbol{\beta}) = f(\mathbf{z}_k{}^T\boldsymbol{\beta})(1 - f(\mathbf{z}_k{}^T\boldsymbol{\beta}))$; and for ordinary Poisson regression: $f'(\mathbf{z}_k{}^T\boldsymbol{\beta}) = f(\mathbf{z}_k{}^T\boldsymbol{\beta})$. Thus, for all three: $v_k \propto f'(\mathbf{z}_k{}^T\boldsymbol{\beta})$. This is not the case for generalized least squares linear regression, however, where the v_k vary across the elements of the population or the ε_k are correlated in some manner.

If $\mathrm{E}(\varepsilon_k{}^2 | \mathbf{z}_k) \propto v(\mathbf{z}_k) < \infty$, and $\mathrm{E}(\varepsilon_k\, \varepsilon_i\, | \mathbf{z}_k, \mathbf{z}_i) = 0$ for $k \neq i$, then the pseudo-ML estimator $\mathbf{b}$ in Equation (12) is consistent under the standard model. When the standard model holds and the analysis weights are ignorable in expectation, however, a more efficient estimator for the model parameter $\boldsymbol{\beta}$ is the solution to $\sum_S \left(f'(\mathbf{z}_k{}^T \mathbf{b})/v_k \right) \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T \mathbf{b}) \right] = \mathbf{0}$.

When the standard model holds, Pfeffermann and Sverchkov (1999) point out that if, in addition, the weights are not ignorable in expectation, $\mathrm{E}(\varepsilon_k{}^2 | \mathbf{z}_k) = v_k < \infty$, and $\mathrm{E}(\varepsilon_k\, \varepsilon_i\, | \mathbf{z}_k, \mathbf{z}_i) = 0$ for $k \neq i$ then a more efficient estimator than the solution to Equation (8) would factor each weight w_k in Equation (11) by $1/\omega(\mathbf{z}_k)$ where $\omega(\mathbf{z}_k)$ is an approximation for $w_k v_k / f'(\mathbf{z}_k{}^T \mathbf{b})$ when v_k is known (up to a constant); otherwise v_k can be replaced by $e_k{}^2 = [y_k - f(\mathbf{z}_k{}^T\mathbf{b})]^2$.

A possible way of generating $\omega(\mathbf{z}_k)$ involves an unweighted Poisson regression of $w_k e_k{}^2$ (or w_k when $v_k \propto h(\mathbf{z}_k)$ is assumed) on the components of $\mathbf{z}_k = (z_{1k}, \ldots, \mathbf{z}_{Pk})^T$ and, perhaps, functions of those components (e.g., $\log(z_{1k})$). Poisson regression is recommended because $w_k e_k{}^2$ (and w_k) is always positive. Recall that in Poisson regression $\log(w_k v_k)$ (or $\log(w_k)$) is modeled as a linear function of components or functions of components.

When the standard model holds in the population, and $\mathrm{E}(\varepsilon_k\, \varepsilon_i\, | \mathbf{z}_k, \mathbf{z}_i) \approx 0$ for $k \neq i$ can be assumed, one can try the following:

1. Fit the estimating equation in (8) and compute the $e_k = y_k - f(\mathbf{z}_k{}^T\mathbf{b})$ (this step is unnecessary when v_k can be assumed to be proportional to $h(\mathbf{z}_k)$ for a known $h(.)$).
2. Fit $w_k e_k{}^2$ (or w_k when $v_k \propto h(\mathbf{z}_k)$ is assumed) on functions of the components of $\mathbf{z}_k$ using unweighted Poisson regression. Call the fitted value $\omega_v(\mathbf{z}_k)$ (when fitting w_k, call the fitted value $\psi_v(\mathbf{z}_k)$, then call $\omega_v(\mathbf{z}_k)$ the product of $\psi_v(\mathbf{z}_k)$ and $h(\mathbf{z}_k)$). Set $\omega_k = \omega_v(\mathbf{z}_k)/f'(\mathbf{z}_k{}^T\mathbf{b})$.
3. Adjust each w_k in the estimating equation in (8) by multiplying it by $1/\omega_k$.
4. Refit the estimating equation in (8) with the w_k replaced by the adjusted weights from step 3 (i.e., w_k/ω_k)).

When the fit in step 2 is good, these steps should return more efficient estimators for the components of $\boldsymbol{\beta}$ than fitting Equation (8) and stopping. We will call this three- or four-step process or any variant of it (e.g., one using linear rather than Poisson regression in step 2) a *P-S weight adjustment*.

 https://doi.org/10.3768/rtipress.2022.mr.0047.2203

Variance Estimation Via Linearization

We restrict attention for now to stratified or single-stratum probability samples of primary sampling units (PSUs) of fixed size without unit nonresponse or coverage error. Additional stages of probability samples can be conducted independently within each PSU to draw the sample elements. We do not rule out samples of elements where the PSUs are completely enumerated or where each PSU is composed of a single element.

In our asymptotic framework, the number of sampled PSUs grows infinitely large along with the population. The number of strata may also grow infinitely large. If so, the number of PSUs in each stratum is assumed to be bounded. Alternatively, the number of strata can be fixed while the number of PSUs in each grows infinitely large. Scenarios where the number of strata grows large but not as fast as the number of sampled PSUs are also possible, but they are not explicitly treated here.

Whether the number of strata should be treated as fixed in an asymptotic framework depends on the design. For example, a design with 60 strata containing two sampled PSUs in each is more reasonably treated in an asymptotic framework where the number of strata grows large, whereas a design with four strata each having over 15 sampled PSUs is more reasonably treated in an asymptotic framework with a fixed number of strata.

Let h denote one of H strata, $\mathbf{u}_k = (u_{k1}, \ldots, u_{kH})^T$ the H-vector of stratum-inclusion identifiers for element k (i.e., $u_{kh} = 1$ when k is in stratum h, and 0 otherwise). Let N (n) denote the number of PSUs in the population (sample), N_h (n_h) the number of PSUs in the population (sample) and stratum h, M (m) the number of elements in the population (sample), M_{hj} (m_{hj}) the number of elements in the population (sample) and PSU j of stratum h, and S_{hj} the set of m_{hj} elements in PSU j of stratum h. We assume that there is a B_M such that in every population in the sequence of populations:

$$M_{hj} \leq B_M < \infty \text{ for all } hj. \tag{13}$$

When First-Stage Stratification Is Ignorable in Expectation

Variance estimation given a stratified multistage sample can be tricky unless a simplifying assumption is made. Usually, the assumption is that the PSUs are randomly selected *with replacement* within strata.

We can instead make the following two ignorability assumptions about the stratum identifiers under the standard (extended) model when reasonable:

1. $E(\varepsilon_k|\mathbf{z}_k,\mathbf{u}_k) = 0$ ($E(\mathbf{z}_k\varepsilon_k|\mathbf{u}_k) = \mathbf{0}$ for the extended model); that is, the first-stage stratification is ignorable in expectation.
2. $E(\varepsilon_k\varepsilon_j\,|\mathbf{z}_k,\mathbf{u}_k,\mathbf{z}_j,\mathbf{u}_j) = 0$ ($(E(\mathbf{z}_k\varepsilon_k\,\mathbf{z}_j\varepsilon_j\,|\mathbf{u}_k,\,\mathbf{u}_j) = 0$ for the extended model) when k and j are from different PSUs and is bounded otherwise.

Although it is likely that strata are chosen such that the mean of the y_k differed across strata, it is nonetheless reasonable to assume that the $E(\varepsilon_k|\mathbf{z}_k)$ (or $E(\mathbf{z}_k\varepsilon_k)$) are unaffected by the first-stage stratum identifiers especially because $\mathbf{z}_k$ in Equation (1) may contain a bounded number (as the number of PSUs grows arbitrarily large) of stratum identifiers or functions of stratum identifiers (e.g., $u_{kh}z_{kp}$).

To estimate the variance of the consistent estimator $\mathbf{b}$ for $\boldsymbol{\beta}$, one starts with this variation of Equation (9),

$$\mathbf{b} - \boldsymbol{\beta} = \left[\sum_{k\in S} w_k f'(\theta_k)\,\mathbf{z}_k\mathbf{z}_k{}^T\right]^{-1} \sum_{k\in S} w_k\mathbf{z}_k\varepsilon_k, \tag{14}$$

for some θ_k between $\mathbf{z}_k{}^T\mathbf{b}$ and $\mathbf{z}_k{}^T\boldsymbol{\beta}$ and the (previously made) assumption that $\mathbf{A}_\theta = M^{-1}\sum_S w_k f'(\theta_k)\mathbf{z}_k\mathbf{z}_k{}^T$ and its probability limit, $\mathbf{A}^\star$, have finite components and are positive definite. For now, we are assuming that the analysis weights, w_k , equal $1/\pi_k$ in this discussion of variance estimation under the extended model. For the standard model, the $1/\pi_k$ can be scaled by a function of the components of $\mathbf{z}_k$.

From Equation (14), we can see the bias of $\mathbf{b}$ is nearly 0. Consequently, a good estimator for its mean-squared-error is also a good estimator for its variance.

As long as all $n_h \geq 2$, the design-based variance/mean-squared-error estimator for $\mathbf{b}$ (from Binder, 1983) is

$$\mathbf{var}(\mathbf{b}) = \mathbf{D}\sum_{h=1}^{H}\frac{n_h}{n_h-1} \times \sum_{j=1}^{n_h}\left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k - \frac{1}{n_h}\sum_{a=1}^{n_h}\sum_{\kappa\in S_{ha}} w_\kappa\mathbf{z}_\kappa e_\kappa\right) \times \left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k - \frac{1}{n_h}\sum_{a=1}^{n_h}\sum_{\kappa\in S_{ha}} w_\kappa\mathbf{z}_\kappa e_\kappa\right)^T\mathbf{D} \quad (15)$$

$$= \mathbf{D}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\left[\sum_{j=1}^{n_h}\left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k\right)\left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k\right)^T - \frac{1}{n_h}\left(\sum_{a=1}^{n_h}\sum_{\kappa\in S_{ha}} w_\kappa\mathbf{z}_\kappa e_\kappa\right)\left(\sum_{a=1}^{n_h}\sum_{\kappa\in S_{ha}} w_\kappa\mathbf{z}_\kappa e_\kappa\right)^T\right]\mathbf{D}$$

where $\mathbf{D} = \left[\sum_S w_k f'(\mathbf{z}_k{}^T\mathbf{b})\,\mathbf{z}_k\mathbf{z}_k{}^T\right]^{-1}$ estimates $M^{-1}\mathbf{A}_\theta^{-1}$ (see Equation (10)), and $e_k = y_k - f(\mathbf{z}_k{}^T\mathbf{b})$.

This is often called the (Taylor-series) *linearization estimator* because, among other things, **D** is a linearized approximation of $(\mathrm{M}\mathbf{A}_\theta)^{-1}$.

Our assumptions assure the near unbiasedness of the variance estimator in equation (15) (as n grows arbitrarily large) given a sampling design and a population such $p\lim(n\Psi_q{}^2)$ is bounded, where Ψ_q is defined in the equation after equation (7). They also assure the nearly unbiasedness of

$$\mathbf{var}_A(\mathbf{b}) = \mathbf{D}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k\right)\left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k\right)^T\mathbf{D}. \quad (16)$$

From a model-based viewpoint, the keys to both variance estimators are (1) the expressions

$$\mathbf{E}_{hj}{}^{\varepsilon} = \sum_{k\in S_{hj}} w_k\mathbf{z}_k\varepsilon_k \quad (17)$$

in Equation (16) have mean **0** and are uncorrelated across PSUs, and (2) $\mathbf{A}^\star$ is the probability limit of $M^{-1}\mathbf{D}^{-1}$. The use of robust sandwich-type variance estimates like Equations (15) and (16) (the **D** being the bread of the sandwich) allows the variance matrices of the $\mathbf{E}_{hj}{}^{\varepsilon}$ to be unspecified. Mild additional asymptotic assumptions allow $\mathbf{E}_{hj} = \sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k$ with $e_k = y_k - \mathbf{z}_k{}^T\mathbf{b} = \varepsilon_k - \mathbf{z}_k{}^T(\mathbf{b}-\boldsymbol{\beta})$ to be used in place of its near equal $\mathbf{E}_{hj}{}^{\varepsilon}$ and $M^{-1}\mathbf{D}$ to replace its near equal $\mathbf{A}_\theta$.

Additional variations of the variance/mean-squared-error estimator in Equation (15) can be made if the analyst is willing to assume that the ε_k are uncorrelated across secondary sampling units or across elements.

The more components there are in $\mathbf{z}_k$, the more reasonable the assumption that the ε_k are uncorrelated across elements (or another higher-stage of sampling like housing units in a household-based sample of individuals) and the more reasonable the assumption that the first-stage stratification is ignorable.

When First-Stage Stratification Is Not Ignorable

Suppose the first-stage stratification is not ignorable and again (for simplicity) $w_k = 1/\pi_k$. Under probability-sampling theory, the $\mathbf{E}_{hj}{}^{\varepsilon}$ could be uncorrelated and have a common mean within strata if the first-stage PSUs had been selected with replacement. That would have allowed the same PSU to be selected more than once, with each selection treated as independent with independent subsampling of elements. Equation (15) (but not Equation (16)) provides a nearly unbiased variance estimator for **b** under such a design. Under many probability-sampling designs employing without-replacement sampling of a fixed number of PSUs, Equation (15) provides, if anything, a slight overestimation of the variances of the components of **b** (which are the diagonals of **var**(**b**)). We will assume our PSU sample has been drawn in such a manner and that the resulting bias in Equation (15) is small enough to be ignored in practice.

Graubard and Korn (2002) point out that when the number of (first-stage) strata remains the same as the population grows arbitrarily large, then Equation (15) provides a nearly unbiased variance estimator under the with-replacement sampling of PSUs only when the fraction of the element population in each stratum is fixed. Otherwise, the fraction of the population within each stratum is a component of the variance of **b** that Equation (15) fails to capture. To avoid that problem, we assume that the fraction of PSUs and elements within each stratum is fixed as the population grows arbitrarily large.

Observe that the variance estimator in Equation (15) can be rewritten as

$$\mathbf{var}(\mathbf{b}) = \mathbf{var}_A(\mathbf{b}) - \mathbf{D}\left(\sum_{h=1}^{H}\frac{1}{(n_h-1)}\sum_{j=1}^{n_h}\sum_{\substack{a=1\\a\neq j}}^{n_h}\mathbf{E}_{hj}\mathbf{E}_{ha}{}^T\right)\mathbf{D}.$$

If the $\mathbf{E}_{hj} \approx \mathbf{E}_{hj}{}^{\varepsilon}$ within each stratum h have a common mean, then the expected values of the diagonals of

var(**b**) (the estimated variances of the components of **b**) will tend to be no higher than the expected values of the diagonals of $\mathbf{var}_A(\mathbf{b})$. They will tend to be lower when some of the stratum means are nonzero. That is, the diagonals of $\mathbf{var}_A(\mathbf{b})$, if anything, tend to overestimate the variances of the components of **b**.

From the above expression we can see that practice of collapsing "similar" strata into variance strata for variance estimation purposes (using Equation (15) with the h indexing the variance strata rather than the design strata) can only bias variance estimation upward. How much upward bias depends on how dissimilar the expectations of the $\mathbf{E}_{hj}$ across the design strata being collapsed into a variance stratum. One popular complex sampling design selects a single PSU per design stratum and collapses pars of "adjacent" (in some sense) design strata into variance strata because Equation (15) requires each n_h to be at least 2.

When every PSU in a design stratum is selected into the sample, these certainty PSUs become the variance strata for use in Equation (15) and the units chosen from them in the next stage of sampling (e.g., housing units selected from area clusters) are variance PSUs.

Calibrated Weight Adjustment

Let d_k be the inverse of the probability that sampled element k has randomly selected for a stratified multistage sample before any weight adjustments for unit nonresponse, frame incompleteness, or efficiency improvement; $d_k = 0$ when $k \in U$ is not a sampled element. The value $q_k = w_k/d_k$ for sampled k is the product of possibly multiple calibration factors ($q_k = 0$ otherwise). There can be multiple adjustments for nonresponse in a complex survey because nonresponse can occur at various levels (e.g., at the household and at the individual).

To simplify the exposition, we will assume that there is a single calibration factor of the form $q_k = S_k q(\mathbf{x}_k{}^T\mathbf{g})$, where $q(t)$ is a monotonic function, such as $q(t) = 1 + \exp(t)$; $\mathbf{x}_k$ is a vector of variables with a finite number of components, including 1 or the equivalent; S_k is 1 when k is in the respondent sample, 0 otherwise; and **g** (if it exists given the range restriction on $q(t)$) satisfies the following calibration equation (observe that a term in either of the two summations below is 0 when k is not in the sample; consequently, $\sum_{k\in S}$ could have been used in place of either $\sum_{k\in U}$ without changing the equation's meaning):

$$\sum_{k\in U} w_k \mathbf{c}_k = \sum_{k\in U} d_k S_k q(\mathbf{x}_k{}^T\mathbf{g})\,\mathbf{c}_k = \mathbf{T}_\mathbf{c}, \qquad (18)$$

where $\mathbf{c}_k$ is a vector of calibration variables with the same number of components as $\mathbf{x}_k$ (the range restrictions on $q(t)$ may render Equation (18) unsolvable for **g**). The population total of $\mathbf{c}_k$—or a nearly unbiased estimate of that total— is known and denoted as $\mathbf{T}_\mathbf{c}$. In practice, the $\mathbf{c}_k$ and $\mathbf{x}_k$ are often but not always identical (with the group-ratio model in Equation (4), $\mathbf{x}_k = \boldsymbol{\delta}_k$ while $\mathbf{c}_k = \boldsymbol{\delta}_k q_k$).

When Equation (18) is used to create calibrated weights that account for unit nonresponse, the components of $\mathbf{T}_\mathbf{c}$ can be estimates from the sample before unit nonresponse; that is, $\mathbf{T}_\mathbf{c} = \sum_U d_k c_k$. The probability of (unit) response is assumed to have the form $1/q(\mathbf{x}_k{}^T\boldsymbol{\gamma})$, and the **g** that satisfies Equation (1) is a consistent estimator of $\boldsymbol{\gamma}$. For example, if response is assumed to be a logistic function of $\mathbf{x}_k$, then $q_k \approx 1+ \exp(\mathbf{x}_k{}^T\boldsymbol{\gamma})$.

We further assume that when an element is sampled, its probability of response is Poisson, that is, independent across the elements of the population. The respondent sample can be treated as a stratified multistage sample.

When Equation (18) is used to calibrate weights that account for coverage error, $1/q(\mathbf{x}_k{}^T\mathbf{g})$ estimates the expected number of times element k is in the sampling frame ($1/q(\mathbf{x}_k{}^T\boldsymbol{\gamma})$). This value can exceed 1 when there is duplication in the frame. More often, the frame is incomplete, and $1/q(\mathbf{x}_k{}^T\boldsymbol{\gamma})$ lies between 0 and 1. Here, we will assume duplication does not occur in the frame for simplicity. In addition, the number of times k is in the sampling frame (0 or 1) is independent across population elements. Consequently, the sample can still be treated as stratified multistage for variance estimation purposes.

Both the models for response and frame undercoverage are selection models, either representing the self-selection of an element into the respondent sample or the "selection" of an element in the population into the sampling frame.

In the remainder of this section, we limit the discussion to response selection models for convenience.

Kott (2015) points out that when the calibration factor is not used for selection modeling but to increase the efficiency of estimated means and totals $q(t)$ is often set at $1 + t$ (this is linear calibration), exp(t) (raking), or $1/(1 + t)$ (pseudo-empirical likelihood) and $\boldsymbol{\gamma} = \mathbf{0}$. Linear calibration and raking are often also used for unit nonresponse adjustment, but then $\boldsymbol{\gamma}$ is no longer $\mathbf{0}$. For unit nonresponse adjustment, setting $q(t) = [L+ \exp(t)]/[1 + \exp(t)/U]$ assumes response is a bounded logistic (or logit) function with response probabilities between $1/U$ and $1/L$.

Let us assume for now that the Poisson selection model for response implied by $q(\mathbf{x}_k{}^T\boldsymbol{\gamma})$ is correct. In addition, when $\mathbf{T_c}$ is a random variable, assume it is uncorrelated with whether element k is a respondent when sampled. Under mild conditions, paralleling the conditions used to justify both Equation (9) and the consistency of $\mathbf{b}$ and (again) invoking the mean-value theorem:

$$\begin{aligned}\mathbf{g} - \boldsymbol{\gamma} \;=\; & \Big[M^{-1} \sum_{k\in U} d_k S_k q'(\varphi_k)\, \mathbf{c}_k \mathbf{x}_k{}^T\Big]^{-1} \\ & M^{-1}\Big[\mathbf{T_c} - \sum_{k\in U} d_k S_k q(\mathbf{x}_k{}^T\boldsymbol{\gamma})\, \mathbf{c}_k\Big]\end{aligned}$$

for some φ_k between $\mathbf{x}_k{}^T\mathbf{g}$ and $\mathbf{x}_k{}^T\boldsymbol{\gamma}$. As a result, $\mathbf{g}$ is a consistent estimator for $\boldsymbol{\gamma}$.

Returning to the model in Equation (1) we are trying to fit, many of the components of $\mathbf{x}_k$ will also often be components of $\mathbf{z}_k$. If they all were or if we replace the standard model assumption in Equation (2) by

$$\mathrm{E}(\varepsilon_k|\, \mathbf{z}_k, \mathbf{x}_k) = 0 \text{ for all realized } \mathbf{z}_k \text{ and } \mathbf{x}_k,\ k\in U, \quad (19)$$

then it is easy to see from

$$\begin{aligned}\mathbf{b} - \boldsymbol{\beta} &= \Big[M^{-1} \sum_{k\in U} w_k f'(\theta_k)\, \mathbf{z}_k \mathbf{z}_k{}^T\Big]^{-1} M^{-1} \sum_{k\in U} w_k \mathbf{z}_k \varepsilon_k \quad (20)\\ &\approx \Big[M^{-1} \sum_{k\in U} d_k S_k q(\mathbf{x}_k{}^T\mathbf{g}) f'(\mathbf{z}_k{}^T\mathbf{b})\, \mathbf{z}_k \mathbf{z}_k{}^T\Big]^{-1} \\ & \quad M^{-1} \sum_{k\in U} d_k S_k q(\mathbf{x}_k{}^T\mathbf{g})\, \mathbf{z}_k \varepsilon_k\end{aligned}$$

that Equation (15) can be used to estimate the variance of $\mathbf{b}$ given $\mathbf{T_c}$. The conditioning on $\mathbf{T_c}$ is needed when $\mathbf{T_c}$ itself is an estimator.

The assumption in Equation (19) collapses to that of the standard model in Equation (2) when the components of $\mathbf{x}_k$ are also in $\mathbf{z}_k$. Under this assumption, we can replace w_k by d_k in defining $\mathbf{b}$, and the estimator will remain consistent.

When the assumption in Equation (19) fails, $\mathbf{b}$ defined with w_k remains a consistent estimator for $\boldsymbol{\beta}$ under the extended model, but variance estimation is confounded by the random variable $\mathbf{g}$ on the right-hand side of Equation (20). It may be approximately equal to $\boldsymbol{\gamma}$, but the approximation is not close enough to be ignored.

Let us assume that the probability sampled element k responds is $1/\, q(\mathbf{x}_k{}^T\boldsymbol{\gamma})$ and that this probability is independent of whether any other sampled element responds. It is not hard to see that

$$\mathbf{b} - \boldsymbol{\beta} \;\approx\; \mathbf{A}^{\star -1} M^{-1} \sum_{k\in U} d_k S_k q(\mathbf{x}_k{}^T\mathbf{g})\, \mathbf{z}_k \varepsilon_k.$$

Let $\xi_k^\star$ be the p^{th} component of $M^{-1}\mathbf{A}^{\star -1}\mathbf{z}_k\varepsilon_k$, so that the error of the p^{th} component of $\mathbf{b}$ (making use of the mean-value theorem) is

$$\begin{aligned}\sum_{k\in S} w_k \xi_k^\star &= M^{-1} \sum_{k\in U} d_k\{c_k^T\boldsymbol{\delta}^\star + S_k(\xi_k^\star - c_k^T\boldsymbol{\delta}^\star)\, q(\mathbf{x}_k^T\mathbf{g})\} \\ &\approx M^{-1} \sum_{k\in U} d_k\{c_k^T\boldsymbol{\delta}^\star + S_k(\xi_k^\star - c_k^T\boldsymbol{\delta}^\star)\, q(\mathbf{x}_k^T\boldsymbol{\gamma}) \\ &\qquad + S_k(\xi_k - c_k^T\boldsymbol{\delta}^\star)\, q'(\mathbf{x}_k^T\boldsymbol{\gamma})\, \mathbf{x}_k^T(\mathbf{g} - \boldsymbol{\gamma})\} \\ &\approx M^{-1} \sum_{k\in U} d_k\{c_k^T\boldsymbol{\delta}^\star + S_k(\xi_k^\star - c_k^T\boldsymbol{\delta}^\star)\, q(\mathbf{x}_k^T\boldsymbol{\gamma})\}, \quad (21)\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\delta}^\star = p\lim\Big\{ & Z\Big(\sum_{j\in U} d_j S_j q'(\mathbf{x}_j^T\mathbf{g})\, \mathbf{x}_j \mathbf{c}_j^T\Big)^{-1} \\ & \sum_{j\in U} d_j S_j q'(\mathbf{x}_j^T\mathbf{g})\, \mathbf{x}_j \xi_j\Big\} \quad (22)\end{aligned}$$

is only needed for dropping the $M^{-1}\sum_U d_k\{S_k(\xi_k - c_k^T\boldsymbol{\delta}^\star)\, q'(\mathbf{x}_k^T\boldsymbol{\gamma})\, \mathbf{x}_k^T(\mathbf{g} - \boldsymbol{\gamma})\}$ term, which also requires asymptotic theory. Both $\mathbf{x}_k{}^T(\mathbf{g} - \boldsymbol{\gamma})$ and $M^{-1}\sum_U d_k\{S_k(\xi_k^\star - c_k^T\boldsymbol{\delta}^\star)\, q'(\mathbf{x}_k^T\boldsymbol{\gamma})\}$ are $\mathrm{O}_p(1/\sqrt{n})$ under mild conditions, so their product is $\mathrm{O}_p(1/n)$, which is small enough to be ignored.

Because the probability of response is Poisson, we can treat the sample as a stratified multistage design, with $\tilde{\xi}_k^\star \;=\; M^{-1}\{c_k^T\boldsymbol{\delta}^\star + S_k(\xi_k^\star - c_k^T\boldsymbol{\delta}^\star)\, q(\mathbf{x}_k^T\boldsymbol{\gamma})\}$ as the element values in an asymptotically equivalent expression for the error of the p^{th} coefficient of $\mathbf{b}$: $\sum_S d_k \tilde{\xi}_k^\star$. The linearized version of this, which is what we would actually use in the variance

estimator, is $\sum_S d_k \tilde{\xi}_k$, with $\tilde{\xi}_k = M^{-1}\{c_k^T \boldsymbol{\delta} + S_k(\xi_k - c_k^T \boldsymbol{\delta})\, q(\mathbf{x}_k^T \mathbf{g})\}$ replacing the p^{th} component of $\mathbf{D}\mathbf{z}_k e_k$ in Equation (15) or (16), ξ_k being the p^{th} component of $M^{-1}\mathbf{A}^{\star -1}\mathbf{z}_k(y_k - \mathbf{z}_k{}^T\mathbf{b})$, and $\boldsymbol{\delta} = \left(\sum_U d_j S_j q'(\mathbf{x}_j^T \mathbf{g})\, \mathbf{x}_j \mathbf{c}_j^T\right)^{-1} \sum_U d_j S_j q'(\mathbf{x}_j^T \mathbf{g})\, \mathbf{x}_j \xi_j$.

With replication, we do an asymptotically equivalent exercise but without having to compute some of the complicated terms like ξ_k and $\boldsymbol{\delta}$ when the standard model fails. Instead, replicate versions of $\mathbf{g}$ are computed in replicate calibration Equations (18). In the next section, we explore one such replication technique: the jackknife.

When calibrating to a constant $\mathbf{T_c}$ in Equation (18) (or, more appropriately, both sides of Equation (18) divided by M, where $M^{-1}\mathbf{T_c}$ remains constant as the population size grows), $\tilde{\xi}_k$ becomes $M^{-1}\{S_k(\xi_k - c_k^T \boldsymbol{\delta}^\star)\, q(\mathbf{x}_k^T \boldsymbol{\gamma})\}$ because $\sum_U d_k c_k^T \boldsymbol{\delta}^\star$ is replaced by a constant $\mathbf{T_c}\boldsymbol{\delta}^\star$, which does not contribute to the variance. Moreover, some of the components of $\mathbf{T_c}$ can come from the full sample whereas some components can be constants or provided from outside samples.

Jackknife Variance Estimation

Replication techniques provide alternative methods for estimating the variance of $\mathbf{b}$ that are especially useful when fitting the extended regression model with calibrated weights. Here, we focus on two forms of jackknife variance estimation, starting with the popular delete-1 (PSU) jackknife. Although often considered a technique for estimating variances under probability-sampling theory (as in Rust, 1985), delete-1 jackknife can also be viewed as a variance estimator under a robust model (Wu, 1986).

Redefine S_{hj} slightly as the set of all respondents in variance PSU j and stratum h, and let S_{h+} be all respondents in variance stratum h. We define the hj^{th} jackknife replicate of $\mathbf{b}$ as the solution ($\mathbf{b}^{(hj)}$) to

$$\sum_{k \in S} w_k{}^{(hj)} \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T \mathbf{b}^{(hj)})\right] = 0, \qquad (23)$$

where $w_k{}^{(hj)} = 0$ when $k \in S_{hj}$,

$w_k{}^{(hj)} = [n_h/(n_h - 1)]\, w_h$ when $k \in S_{h+}$ but $k \notin S_{hj}$,

and $w_k{}^{(hj)} = w_k$ otherwise.

This is identical to the estimator $\mathbf{b}$ (for $\boldsymbol{\beta}$) in Equation (8) computed from a sample paralleling S except that only $n_h - 1$ variance PSUs from stratum h are included, that is, all the variance PSUs in h except hj. Consequently, analogous to Equations (9) and (10)

$$\mathbf{b}^{(hj)} - \boldsymbol{\beta} = \left[\mathbf{A}_\theta^{(hj)}\right]^{-1} M^{-1} \sum_{k \in S} w_k{}^{(hj)} \mathbf{z}_k \varepsilon_k, \qquad (24)$$

where $\mathbf{A}_\theta^{(hj)} = M^{-1} \sum_{k \in S} w_k{}^{(hj)} f'(\theta_k)\, \mathbf{z}_k \mathbf{z}_k{}^T$.

The limit of $\mathbf{A}_\theta^{(hj)}$ as the number of sampled PSUs gets arbitrarily large is $\mathrm{A}^\star$, just like $\mathbf{A}_\theta$. Consequently,

$$\begin{aligned}
\mathbf{b}^{(hj)} - \mathbf{b} &\approx [\mathbf{A}^\star]^{-1} M^{-1} \left(\sum_{k \in S} w_k{}^{(hj)} \mathbf{z}_k \varepsilon_k - \sum_{k \in S} w_k \mathbf{z}_k \varepsilon_k \right) \\
&= [\mathbf{A}^\star]^{-1} M^{-1} \left(\left[\frac{1}{n_h - 1} \sum_{\substack{k \in S_{h+} \\ k \notin S_{hj}}} w_k \mathbf{z}_k \varepsilon_k \right] - \sum_{k \in S_{hj}} w_k \mathbf{z}_k \varepsilon_k \right) \\
&= [\mathbf{A}^\star]^{-1} M^{-1} \frac{n_h}{n_h - 1} \left(\left[\frac{1}{n_h} \sum_{k \in S_{h+}} w_k \mathbf{z}_k \varepsilon_k \right] - \sum_{k \in S_{hj}} w_k \mathbf{z}_k \varepsilon_k \right),
\end{aligned}$$

and some more algebra reveals that the delete-1 jackknife variance estimator for $\mathbf{b}$,

$$\mathbf{var}_{\mathrm{D1J}}(\mathbf{b}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\mathbf{b}^{(hj)} - \mathbf{b})\, (\mathbf{b}^{(hj)} - \mathbf{b})^T, \qquad (25)$$

is nearly equal to the (Taylor-series) linearization variance estimator in Equation (15).

There are two main differences between $\mathbf{var}(\mathbf{b})$ in Equation (15) and $\mathbf{var}_{\mathrm{D1J}}(\mathbf{b})$ in Equation (25). The former replaces the ε_k with e_k. This often causes $\mathbf{var}(\mathbf{b})$ to slightly underestimate the variances of the components of $\mathbf{b}$ when the number of sampled PSUs is not "arbitrarily large," such as in actual finite-world application. The delete-1 jackknife does not make that replacement. Instead, it treats $\mathbf{A}_\theta^{(hj)}$ as if it were the same as $\mathbf{A}_\theta$. This often causes $\mathbf{var}_{\mathrm{D1J}}(\mathbf{b})$ to slightly overestimate the variances of the components of $\mathbf{b}$ when the number of sampled PSUs is not arbitrarily large.

The delete-1 jackknife produces as many sets of jackknife replicate weights as there are variance PSUs. Many find handling so many sets of weights (including the original weights) burdensome. When $n_h = 2$ in every variance stratum, an alternative delete-1 jackknife creates replicate weights for only one variance PSU per variance stratum and computes

$$\mathbf{var}_{\mathrm{D1J\text{-}alt}}(\mathbf{b}) = \sum_{h=1}^{H} (\mathbf{b}^{(h1)} - \mathbf{b})\, (\mathbf{b}^{(h1)} - \mathbf{b})^T. \qquad (26)$$

https://doi.org/10.3768/rtipress.2022.mr.0047.2203

The Delete-a-Group Jackknife

Several packages can compute other jackknife variance estimates with a reduced number of sets of jackknife replicate weights (one for each replicate). The generic form of the jackknife is

$$\mathbf{var}_{\mathrm{GJ}}(\mathbf{b}) = \sum_{r=1}^{R} M_r(\mathbf{b}^{(r)} - \mathbf{b})(\mathbf{b}^{(r)} - \mathbf{b})^T, \qquad (27)$$

where each $\mathbf{b}^{(r)}$ is computed with its own set of replicate weights. Observe that Equations (25) and (26) have this generic form.

To run a delete-a-group (DAG) jackknife, first sort the variance PSUs by variance stratum and assign each variance PSU systematically to one of R replicate groups, which are not replicates, although there will ultimately be R sets of DAG jackknife replicate weights. In addition, $M_r = (R\text{-}1)/R$ for all r in Equation (27). The number of replicate groups needs to be large enough for the resulting variance estimator to be relatively stable, say $R = 30$.

Let h denote a variance stratum as before, r a replicate group, and S^{hr} the set of sampled respondents in both variance stratum h and replicate group r. Let n_h be the number of sampled PSUs in variance stratum h.

When $n_h \geq R$, the R DAG jackknife replicate weights are computed for each sampled respondent k in variance stratum h, as follows:

$w_k^{(r)} = 0$ when $k \in S^{hr}$, and

$w_k^{(r)} = w_k n_h/(n_h - n_{hr})$ when $k \notin S^{hr}$, which explains the name.

When $n_h < R$, the R DAG jackknife replicate weights for a respondent in stratum h are as follows:

$w_k^{(r)} = w_k$ when S^{hr} is empty,

when:

$w_k^{(r)} = w_k[1 - (n_h - 1)Z]$ when $k \in S^{hr}$,

$w_k^{(r)} = w_k(1 + Z)$ when S^{hr} is not empty and $k \notin S^{hr}$,

where $Z = \sqrt{\{[R/(R-1)][n_h(n_h-1)]^{-1}\}}$.

The proof that the DAG jackknife works can be found in Kott (2001). All replicate variance estimators with a form like Equation (27) may exhibit a slight tendency to have an upward bias (which shrinks to 0 as the number of sampled PSUs grows arbitrarily large) caused by $\mathbf{b}^{(r)}$ in $(\mathbf{b}^{(r)} - \mathbf{b})$ being computed with $\mathbf{A}_\theta^{(r)} = M^{-1}\sum_S w_k{}^{(r)} f'(\theta_k)\mathbf{z}_k\mathbf{z}_k{}^T$ rather than with $\mathbf{A}_\theta$ as is $\mathbf{b}$.

When calibrating weights with Equation (18), suppose some of the components of $\mathbf{T_c}$ come from independently drawn samples originating outside of S, each with the same number of DAG jackknife replicate groups as S, and the remaining components are either constants or come from S before unit nonresponse. A DAG jackknife variance estimator can capture both the variance contributed from the outside sample and from S (and from adjusting for unit nonresponse).

Imputing Missing Item Values with a Regression Model

In this section, we change focus from model fitting to prediction, particularly the prediction needed when imputing for a missing survey value. Most complex sample surveys suffer from item nonresponse. This occurs when a sampled (unit) respondent $k \in S$ provides item values for some survey items but not for others. Suppose all survey respondents provide values for the vector of variables $\mathbf{z}_k{}^A$ but only some provide a value for y_k. To estimate the population total, $T_y = \sum_U y_k$, with an analysis-weighted sample, one can compute

$$t_y = \sum_{k\in S} w_k y_k R_k + \sum_{k\in S} w_k f(\mathbf{z}_k{}^T\mathbf{b})(1 - R_k), \qquad (28)$$

where $R_k = 1$ when k is an item respondent, 0 otherwise, and $\mathbf{z}_k$ is a subset of $\mathbf{z}_k{}^A$. Analogously, for estimating the population mean, T_y/M, one can replace all w_k in Equation (28) by $w_k/\sum_S w_j$. In Equation (28), when $R_k = 0$, the missing y_k is imputed by $y_k{}^I = (f(\mathbf{z}_k{}^T\mathbf{b})$.

Suppose the standard regression model relating y_k to $f(\mathbf{z}_k{}^T\boldsymbol{\beta})$ in Equations (1) and (2) holds and the probability of item response for each unit respondent k is wholly a function of the components of $\mathbf{z}_k$. Solving

$$\sum_{k\in S} w_k \Phi(\mathbf{z}_k)\mathbf{z}_k[y_k - f(\mathbf{z}_k{}^T\mathbf{b})]R_k = 0,$$

https://doi.org/10.3768/rtipress.2022.mr.0047.2203

for $\mathbf{b}$, where $\Phi(\mathbf{z}_k)$ is any scalar function of $\mathbf{z}_k$, and plugging that $\mathbf{b}$ into Equation (28) renders t_y in Equation (28) a nearly unbiased estimator for T_y in some sense.

When the standard model does not hold or the probability of item response is not wholly a function of the components of $\mathbf{z}_k$, we can alternatively attempt to find a $\mathbf{b}$ satisfying

$$\sum_{k\in S} w_k \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T\mathbf{b})\right](1 - R_k) = \mathbf{0},$$

We are restricted for computational purposes to item-responding members of S. Consequently, we can try to find a $\mathbf{b}$ satisfying

$$\sum_{k\in S} w_k \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T\mathbf{b})\right](1 - \mathrm{E}(R_k \big| y_k, \mathbf{z}_k{}^A)) \frac{R_k}{\mathrm{E}(R_k|y_k, \mathbf{z}_k{}^A)} = \mathbf{0}$$

or

$$\sum_{k\in S} w_k r_k \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T\mathbf{b})\right] = \mathbf{0}, \qquad (29)$$

where $w_k = \frac{I_k}{\mathrm{E}(I_k|\cdot)}$ is the analysis weight, $r_k = \frac{1 - \mathrm{E}(R_k|y_k, \mathbf{z}_k{}^A)}{\mathrm{E}(R_k|y_k, \mathbf{z}_k{}^A)} R_k$ is the item-response weight, $I_k = 1$ when $k \in S$ (0 otherwise), and $|\cdot$ denotes conditioning on all the variables used in determining the probability of inclusion in respondent sample S. This assumes we have fit an item-response model for R_k. We will describe a method for assuming and fitting such a model later.

In this section, we always assume that $\mathrm{E}(I_k|\cdot)$ is correctly specified and consistently estimated (recall that the analysis weights can include adjustments to compensate for unit nonresponse and frame undercoverage). In addition, if the item-response model $\mathrm{E}(R_k|y_k, \mathbf{z}_k{}^A)$ is correctly specified, then satisfying

$$\sum_{k\in S} w_k \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T\mathbf{b})\right](1 - \mathrm{E}(R_k \big| y_k, \mathbf{z}_k{}^A)) \frac{R_k}{\mathrm{E}(R_k|y_k, \mathbf{z}_k{}^A)} = \mathbf{0},$$

implies

$$\mathrm{E}_R\left[\sum_{k\in U} w_k y_k (1 - R_k) - \sum_{k\in U} w_k f(\mathbf{z}_k{}^T\mathbf{b})(1 - R_k)\right] = 0.$$

As a result, Equation (28) provides a nearly unbiased estimated for T_y in some sense regardless of whether the standard regression model holds as long as $\mathbf{z}_k$ contains 1 or the equivalent because $\sum_S w_k r_k \left[y_k - f(\mathbf{z}_k{}^T\mathbf{b})\right] = 0$ is a component of Equation (29).

An Example

A common example of imputation with a regression model is imputation with the group-ratio model in Equation (4). When q_k varies across the k, $\mathbf{z}_k = q_k\boldsymbol{\delta}_k$ does not contain an intercept as we noted previously. Nevertheless,

$$\begin{aligned} t_y &= \sum_{k\in U} w_k y_k R_k + \sum_{g=1}^{G} \sum_{k\in U} w_k d_{kg} q_k b_g (1 - R_k) \\ &= \sum_{k\in U} w_k y_k R_k + \\ &\quad \sum_{g=1}^{G} \sum_{k\in U} w_k \delta_{kg} q_k (1 - R_k) \frac{\sum_{k\in U} w_k \delta_{kg} y_k \left[1 - \mathrm{E}(R_k|y_k, \mathbf{z}_k{}^A)\right] \frac{R_k}{\mathrm{E}(R_k|y_k, \mathbf{z}_k{}^A)}}{\sum_{k\in U} w_k \delta_{kg} q_k \left[1 - \mathrm{E}(R_k|y_k, \mathbf{z}_k{}^A)\right] \frac{R_k}{\mathrm{E}(R_k|y_k, \mathbf{z}_k{}^A)}} \\ &= \sum_{U} w_k y_k R_k + \sum_{g=1}^{G} \sum_{k\in U} w_k \delta_{kg} q_k (1 - R_k) \frac{\sum_{k\in U} w_k \delta_{kg} y_k R_k}{\sum_{k\in U} w_k \delta_{kg} q_k R_k}. \end{aligned}$$

The last line assumes $\mathrm{E}(R_k \big| y_k, \mathbf{z}_k{}^A)$ is constant within each group.

Observe that t_y above is nearly unbiased in some sense when either of the following is true:

1. The standard group-ratio model holds in the population, the analysis weights are ignorable, and the probability of item nonresponse is wholly a function of $\mathbf{z}_k$ (combined with $\mathrm{E}(w_k|\cdot) = 1$); or
2. The probabilities of item response are constant within each group (and $\mathrm{E}(w_k|\cdot) = 1$).

This property has been called "double robustness," but double protection against item nonresponse bias is a more accurate description.

The leaves of a decision tree (classification or regression) for y_k is a group-mean outcome model. Note that the tree can only be fit among item respondents. Decision tree methodology can be used to fit a group-mean response model. In this case, the entire unit respondent sample can be used to fit the model.

Assuming and Fitting an Item-Response Model

More generally, suppose it is reasonable to assume that the item-response model has the form:

$$\mathrm{E}(R_k|{}_i\mathbf{x}_k) = h({}_i\mathbf{x}_k{}^T{}_i\boldsymbol{\gamma}), \qquad (30)$$

where $h(.)$ is a known function (e.g., $h(\theta) = 1/[1 + \exp(\theta)]$); ${}_i\mathbf{x}_k$ is a vector of survey variables known for all item respondents, which means it may contain

y_k along with components of $\mathbf{z}_k{}^A$ and functions of components of $\mathbf{z}_k{}^A$; and ${}_i\boldsymbol{\gamma}$ is a vector of unknown parameters. The prefix i on ${}_i\mathbf{x}_k$ and ${}_i\boldsymbol{\gamma}$ differentiates them from the vectors in unit response function $1/q(\mathbf{x}_k{}^T\boldsymbol{\gamma})$ described previously.

Let $\mathbf{z}_k{}^0$ be a vector containing components of $\mathbf{z}_k{}^A$ (and functions of components of $\mathbf{z}_k{}^A$) having the same number of components as ${}_i\mathbf{x}_k$. If the item-response model in Equation (30) is correctly specified, then a consistent estimator for ${}_i\boldsymbol{\gamma}$ will be the solution ${}_i\mathbf{g}$ of the calibration equation (if it exists given the range restrictions on h(.))

$$\sum_{k\in U} w_k \mathbf{z}_k{}^0 \left[\frac{1-h({}_i\mathbf{x}_k{}^T{}_i\mathbf{g})}{h({}_i\mathbf{x}_k{}^T{}_i\mathbf{g})}\right] R_k = \sum_{k\in U} w_k \mathbf{z}_k{}^0 [1-R_k],$$

or its near equivalent

$$\sum_{k\in U} w_k \mathbf{z}_k{}^0 \frac{R_k}{h({}_i\mathbf{x}_k{}^T{}_i\mathbf{g})} = \sum_{k\in U} w_k \mathbf{z}_k{}^0. \qquad (31)$$

The number of components in $\mathbf{z}_k{}^0$ is flexible. One can always increase the number of components in $\mathbf{z}_k{}^0$ to equal the number of components in ${}_i\mathbf{x}_k$, thus making the number of implicit equations in (31) (the components of $\mathbf{z}_k{}^0$) equal the number of unknowns in ${}_i\boldsymbol{\gamma}$. If ${}_i\mathbf{x}_k$ has fewer components than $\mathbf{z}_k$, then we can generate $\mathbf{z}_k{}^0$ with

$$\mathbf{z}_k{}^0 = \sum_{k\in S} R_j {}_i\mathbf{x}_j \mathbf{z}_j{}^T \left(\sum_{k\in S} R_j \mathbf{z}_j \mathbf{z}_j{}^T\right)^{-1} \mathbf{z}_k,$$

which essentially regresses the components of ${}_i\mathbf{x}_k$ onto $\mathbf{z}_k$ using ordinary least squares applied to the item respondents.

In many applications, ${}_i\mathbf{x}_k$ in the assumed item-response model (Equation (30)) is equal to $\mathbf{z}_k{}^0$, and $\mathbf{z}_k{}^0$ is made up of components of $\mathbf{z}_k$ and functions of components of $\mathbf{z}_k$. As a result, the solution $\mathbf{b}$ to Equation (29) leads to doubly robust imputation when $\mathbf{z}_k$ contains an intercept (or the equivalent) when either the standard regression model holds and the true (but not specified) item-response model is a function of $\mathbf{z}_k$, or when the assumed item-response model fit using Equation (31) is indeed the true item-response model.

Suppose the standard regression model in Equations (1) and (2) holds. If the model errors (the ε_k in Equation (1)) are uncorrelated and the true item-response model is a function of $\mathbf{z}_k$, then in the spirit of P-S adjustments, we should be able to increase the efficiency of $\mathbf{b}$ by dividing the $w_k r_k$ in Equation (29) by $\omega_k = \omega(\mathbf{z}_k)$, which is $1/f'(\mathbf{z}_k{}^T\mathbf{b})$ times the predicted value of a Poisson regression of $w_k r_k[y_k - f(\mathbf{z}_k{}^T\mathbf{b})]^2$ on appropriately chosen components of $\mathbf{z}_k$ and functions of those components. To obtain double robustness, we may have to add ω_k to the components of $\mathbf{z}_k$ in Equation (1), when it is not already a linear function of those components, to assure that $\sum_S w_k r_k[y_k - f(\mathbf{z}_k{}^T\mathbf{b})] = 0$.

Nonignorable Item Nonresponse

When y_k is a component of ${}_i\mathbf{x}_k$ in the item-response model (that is, item nonresponse is nonignorable), the situation can become more complicated. Fitting Equations (31) and then (29) to determine r_k and then $\mathbf{b}$ will produce a nearly unbiased estimator for T_y when the item-response model in Equation (30) is correctly specified.

If both the item-response model and the standard regression model in Equations (1) and (2) are correctly specified, then $\mathbf{b}$ is a nearly unbiased estimator for $\boldsymbol{\beta}$. This can be softened slightly because of the standard regression model assumption in Equation (2). The estimation of $[1 - \mathrm{E}(R_k|{}_i\mathbf{x}_k)]/\mathrm{E}(R_k|{}_i\mathbf{x}_k)] = [1 - h({}_i\mathbf{x}_k{}^T{}_i\boldsymbol{\gamma})]/h({}_i\mathbf{x}_k{}^T{}_i\boldsymbol{\gamma})$ within $r_k = [1 - \mathrm{E}(R_k|{}_i\mathbf{x}_k)]/\mathrm{E}(R_k|{}_i\mathbf{x}_k)]$ need only be correctly specified up to a function of $\mathbf{z}_k$. Consequently, if we fit $\mathrm{E}(R_k|{}_i\mathbf{x}_k)]$ with $1/[1 + \exp(y_k \mathrm{g}_y + \mathbf{z}_k{}^T\mathbf{g}_\mathbf{z})]$ but the true response function is $1/[1 + \exp(y_k\gamma_y)\varphi(\mathbf{z}_k)]$ for some unknown $\varphi(\mathbf{z}_k)$, and g_y is a consistent estimator for γ_y, then $\mathbf{b}$ remains nearly unbiased under the standard regression model. In practice, g_y may not be a consistent estimator for γ_y after fitting $1/[1 + \exp(y_k \mathrm{g}_y + \mathbf{z}_k{}^T\mathbf{g}_\mathbf{z})]$, and so $\mathbf{b}$ would not be nearly unbiased. Still, $\mathbf{b}$ may be a reasonable, if imperfect, estimator for $\boldsymbol{\beta}$.

We can again potentially increase the efficiency of $\mathbf{b}$ by dividing the $w_k r_k$ in Equation (29) by ω_k as described previously. For estimates of totals and means to remain nearly unbiased under the response model, we may need to add ω_k to the components of $\mathbf{z}_k$ in Equation (1).

Some Concluding Remarks

Summary

Complex surveys are usually designed to estimate population totals, means, and simple ratios of collected survey items. Sometimes, however, analysts want to fit regression models among the items. The population mean of a survey item is the simplest example of a standard regression model, one that always holds in the population but whose consistent estimation can be affected by members of the sample having unequal probabilities of selection. As we have seen, whether the standard model holds and whether unequal selection probabilities affect consistent estimation are two distinct issues.

Given an assumed statistical model, $E(y_k) = f(\mathbf{z}_k{}^T\boldsymbol{\beta})$, relating a survey item y_k for population member k to an explanatory vector of survey items $\mathbf{z}_k$, the standard regression model holds when $E\{[y_k - f(\mathbf{z}_k{}^T\boldsymbol{\beta})]|\mathbf{z}_k\} = 0$ for all realized values of $\mathbf{z}_k$ in the population. That model can, and often does, fail. One reason for its failure is that a complex survey is limited in the variables that can serve as components of $\mathbf{z}_k$. A more reasonable model may require more explanatory variables than available in the survey.

Even when the assumed standard model does not fail, the expectation of the model errors, $\varepsilon_k = y_k - f(\mathbf{z}_k{}^T\boldsymbol{\beta})$, may depend on the elements' probabilities of sample selection. Assuming some mild conditions hold, by injecting the inverses of the element selection probabilities—the analysis weights $\{w_k\}$—into an estimating equation, $\sum_S w_k\mathbf{z}_k\,[y_k - f(\mathbf{z}_k{}^T\mathbf{b})] = \mathbf{0}$ (where S denotes the responding sample) and solving for $\mathbf{b}$, one can consistently estimate $\boldsymbol{\beta}$ under the standard model. Solving this weighted estimating equation for $\mathbf{b}$ also consistently estimates $\boldsymbol{\beta}$ under the more general extended model, which only assumes $E\{\mathbf{z}_k\,[y_k - f(\mathbf{z}_k{}^T\boldsymbol{\beta})]\} = \mathbf{0}$.

An analysis weight w_k can have several factors: the inverse of the probability that element k was randomly selected from the sampling frame, the inverse of the estimated probability that selected element k responded to the survey, the estimated inverse of the probability that population element k was in the sampling frame from which the sample was selected, and a small scaling adjustment to increase the efficiency of estimated item totals. It is important to realize that the second and third components involve estimating a function that can be mis-specified. The first and fourth do not.

If the standard regression model holds, then $\mathbf{b}$ remains a consistent estimator when each analysis weight in the estimating equation is multiplied by a scalar function of the explanatory variables in $\mathbf{z}_k$. That scalar function can be chosen to increase the efficiency of the components of $\mathbf{b}$. In addition, as long as both the true probability of unit response (or frame undercoverage) and the estimate of that probability are both functions of the explanatory variables in $\mathbf{z}_k$, then using the adjusted analysis weights in the weighted estimating equation produces a consistent estimator for $\mathbf{b}$ when the standard regression model holds even when the function used to estimate the unit response probability is mis-specified.

Indeed, when the standard regression model holds, then one need not weight the estimating equation in computing a nearly unbiased $\mathbf{b}$ when the inverse of the probability of selection into the respondent sample is a function of the regression model's explanatory variables. Fitting a standard regression model requires weighting only when the probability of selection into the respondent sample when conditioned on the regression model's explanatory variables is a function of the dependent variable.

Often, with more explanatory variables in $\mathbf{z}_k$, there is less need for analysis weights in the estimating equation. Similarly, with more components in $\mathbf{z}_k$, the standard model is more likely to hold. Kott (2018) discusses tests for assessing whether the standard model holds or whether analysis weights are needed for estimating a regression model.

When estimating a population total or mean with a complex survey, imputing for a missing item value with the predicted value of a regression model using other survey items as the explanatory variables can lead to nearly unbiased estimation in some sense when the standard model holds in the population. In fact, when the standard model holds and item missingness is a function of the explanatory variables and not the item being imputed, it is unnecessary to

use weights when fitting the regression model. Using the products of the analysis weight and an item-response weight when fitting the regression model can protect against the failure of the regression model when the item-response model used for computing the item-response weights is correctly specified and consistently estimated.

We saw that a calibration equation in (31) can be used to fit an item-response model. A calibration equation in (18) can likewise be used to fit a unit response model (or a coverage model) when adjusting analysis weights. When response is partially a function of the dependent variable given the regression model's explanatory variables, these response models need to be correctly specified when the standard regression model fails.

When response is partially a function of the dependent variable, the standard model holds, and the ratio of the true and the fitted but mis-specified response models is a function of the regression model's explanatory variables; using the fitted response model to create the analysis or item-response weight will then produce nearly unbiased estimates. Although this last condition is not likely to be satisfied in practice, it suggests that using a mis-specified response model may remove some potential for bias resulting from nonignorable nonresponse at the unit or item level.

The DAG jackknife provides a useful method for estimating variances of coefficient estimates in a regression model or item means when there is item nonresponse. Linearization is difficult in either case when analysis weights are calibrated. An exception occurs when estimating coefficients under the standard regression model, and the calibration adjustments are function of the explanatory variables in $\mathbf{z}_k$ and perhaps when a vector $\mathbf{x}_k$ such that $E(\varepsilon_k | \mathbf{z}_k, \mathbf{x}_k) = 0$ for all realized $\mathbf{z}_k$ and $\mathbf{x}_k$.

Speculations on Imputation and Variance Estimation

The DAG jackknife can be used to measure the variance of an estimated infinite population mean (the population mean as the population size grows arbitrarily large) computed with Equation (28), where each analysis weight is replaced by $w_k / \sum_S w_j$. With G sets of replicate analysis weights $\{w_{k(g)}, g = 1, \ldots, G\}$ and item-response weights $\{r_{k(g)}, g = 1, \ldots, G\}$, there are likewise G versions of $\mathbf{b}^{(g)}$ and G versions of the imputed value for a missing y_k: $f(\mathbf{z}_k{}^T\mathbf{b})$; namely, $f(\mathbf{z}_k{}^T \mathbf{b}^{(g)})$, $g = 1, \ldots, G$. Each $\mathbf{b}^{(g)}$ is computed with a replicate version of Equation (29). If it exists, an efficiency increasing weighting factor, $1/\omega_k$, need not be replicated.

When a goal is to estimate the distribution of the y_k in the population, the implicit imputation of a missing y_k with $f(\mathbf{z}_k{}^T\mathbf{b})$ in Equation (28) is not helpful. When $f(.)$ is logistic, we can impute a missing y_k with 1 with probability $f(\mathbf{z}_k{}^T\mathbf{b})$ and with 0 otherwise. To determine the probabilities of imputation with 1 in a way that, at most, marginally distorts the estimated mean, sort the m item nonrespondents in random order and assign the first in that order probability $1/(2m)$, so that missing y_k is imputed with 1 when $f(\mathbf{z}_k{}^T\mathbf{b}) > 1/(2m)$ and 0 otherwise. Similarly, assign the second in order probability $3/(2m)$, the third probability $5/(2m)$, …, and the last probability $(2m - 1)/(2m)$. In a DAG jackknife replicate, it is the size of $f(\mathbf{z}_k{}^T\mathbf{b}^{(r)})$ that is compared with $1/(2m)$, …, or $(2m - 1)/(2m)$.

When $f(.)$ is linear, the following seems reasonable: add the residual $y_j - f(\mathbf{z}_j{}^T\mathbf{b})$ from one of the item respondents to $f(\mathbf{z}_k{}^T\mathbf{b})$ when y_k is missing. Similarly, when $f(.)$ is Poisson and the imputed values need to be positive, add $[f(\mathbf{z}_k{}^T\mathbf{b})/ f(\mathbf{z}_j{}^T\mathbf{b})][y_j - f(\mathbf{z}_j{}^T\mathbf{b})]$. To choose which item respondent's residual to use as a donor for element k with a missing item value, first sort the item respondents in size order of the $f(\mathbf{z}_j{}^T\mathbf{b})$ and select a systematic probability proportional to $w_j r_j$ (or $w_j r_j/\omega_j$ if more appropriate) sample of m donors, where m is the number of item nonrespondents and then assign the residuals of the m selected donors to the m item nonrespondents sorted by the size of $f(\mathbf{z}_k{}^T\mathbf{b})$. This matches donors and recipients in some sense while limiting the distortion of the estimated mean caused by adding residuals. In every jackknife replicate, the same donor residual is used when needed for a particular item nonrespondent to avoid overestimating the contribution to variance from adding residuals to the imputation.

 https://doi.org/10.3768/rtipress.2022.mr.0047.2203

For a variable that can be either positive or 0, we can first impute whether the variable is positive using a logistic regression, then if imputed to be positive, impute the positive value with a Poisson regression.

References

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*(3), 279–292. https://doi.org/10.2307/1402588

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382. https://doi.org/10.1080/01621459.1992.10475217

Fuller, W. A. (1975) Regression analysis for sample survey. *Sankhya—The Indian Journal of Statistics*, *37*(Series C), 117–132.

Godambe, V. P., & Thompson, M. E. (1974). Estimating equations in the presence of a nuisance parameter. *Annals of Statistics*, *2*(3), 568–571. https://doi.org/10.1214/aos/1176342718

Graubard, B. I., & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, *17*, 73–96. https://doi.org/10.1214/ss/1023798999

Kott, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, *17*(4), 521–526.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, *32*(2), 133–142.

Kott, P. (2007). Clarifying some issues in the regression analysis of survey data. *Survey Research Methods*, *1*(1), 11–18. https://doi.org/10.18148/srm/2007.v1i1.47

Kott, P. (2015). Calibration weighting in survey sampling. *Wiley Interdisciplinary Reviews: Computational Statistics*, *8*(1), 39–53. https://doi.org/10.1002/wics.1374

Kott, P. (2018). A design-sensitive approach to fitting regression models with complex survey data. *Statistics Surveys*, *12*, 1–17. https://doi.org/10.1214/17-SS118

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley. https://doi.org/10.1002/9781119013563

Pfeffermann, D., & Sverchkov, M. (1999) Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhya—The Indian Journal of Statistics*, *61*(Series B), 166–186.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, *1*, 381–397.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59–87). Wiley.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, *14*(4), 1261–1295. https://doi.org/10.1214/aos/1176350142

RTI International is an independent, nonprofit research institute dedicated to improving the human condition. We combine scientific rigor and technical expertise in social and laboratory sciences, engineering, and international development to deliver solutions to the critical needs of clients worldwide.

www.rti.org/rtipress **RTI Press publication MR-0047-2203**