The Role of Weights in Regression Modeling and Imputation (Including When There is Nonignorable Nonresponse)

Phillip S. Kott
RTI International, 6011 Executive Blvd., Rockville, MD 20852

Abstract
The standard regression model, whether linear, logistic, or Poisson, assumes that the expected value of the model error, the difference between the dependent variable and its model-based prediction, is zero no matter what the values of the explanatory variables. The standard model can fail for a given population. A rarely-failing extended regression model assumes only that the model error is uncorrelated with the model's explanatory variables. Consistent estimates under either the standard or extended model given complex survey data with inverse-probability weights (broadly defined) can be determined by fitting a weighted estimating equation based on the extended model's assumption of the model error being uncorrelated with the 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 item missingness is a function of the explanatory variables of the regression model and not the item value itself (i.e., item values are missing at random), 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. Weights can be adjusted to retain protection against bias when the standard model for the item value fails while increasing the efficiency of the estimated item mean when the standard model for holds among members of the population that would not have provided item values if surveyed.

Key Words: Standard model, Extended model, Item-response model, Inverse-probability weight, Pfeffermann-Sverchkov adjustment.

1. 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 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, for generally).

We do not take that so-called "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 of the methods developed in the conventional design-based framework, such as fitting weighted estimating

equations and sandwich variance/mean-squared-error estimation, are retained but their interpretations change.

Section 2 lays out the design-sensitive approach to regression modeling with complex survey data, which involves drawing a distinction between the robust *standard* model and more general *extended* model. Estimating model parameters under the latter requires the use of inverse-probability weights (broadly defined to include calibration adjustments), while it is possible to adjust those weights under the former to increase the efficiency of parameter estimates. Section 3 describes linearization-based variance estimation for model parameters, while Section 4 explores versions of jackknife replication variance estimation. Section 5 provides some useful tests for determining whether using inverse-probability weights are necessary and whether the standard model holds in the population.

Section 6 changes the focus. This section addresses 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 then extended to situations where item nonresponse is not missing at random, providing an unbiased estimate for an item mean in some sense when the standard model fails and a more efficient estimate when it does not. Finally, Section 7 provides a review of the ideas developed in the preceding sections.

An appendix available from the author upon request provides empirical examples of the some of the methods described in the text using SAS-callable SUDAAN (Research Triangle Institute, 2012).

2. The Design-Sensitive Approach to Regression Modeling

We start by defining the *standard* (regression) *model* in the following distribution-free manner. Given any element (member) k of a population U, the standard model assumes that

$$y_k = f(\mathbf{z}_k^T\boldsymbol{\beta}) + \varepsilon_k, \tag{2.1}$$

where

$$\mathrm{E}(\varepsilon_k \,|\mathbf{z}_k) = 0 \text{ for all realized } \mathbf{z}_k,\ k \in U \tag{2.2}$$

In equation (2.1), y_k is the dependent random variable being modeled, while $\mathbf{z}_k$ is a vector of P explanatory variables (covariates), one of which is 1 (or, equivalently, some linear combination of the components of $\mathbf{z}_k$, is 1), and $f(.)$ is a specified monotonic function. In particular,

$$\begin{aligned} f(\mathbf{z}_k^T\boldsymbol{\beta}) &= \mathbf{z}_k^T\boldsymbol{\beta} && \text{for a linear regression model,} \\ &= \exp(\mathbf{z}_k^T\boldsymbol{\beta})/[1 + \exp(\mathbf{z}_k^T\boldsymbol{\beta})] && \text{for a (simple) logistic regression model,} \\ &= \exp(\mathbf{z}_k^T\boldsymbol{\beta}) && \text{for a Poisson regression model.} \end{aligned}$$

Finally, $\boldsymbol{\beta}$ is a vector of parameter that is unknown but can be estimated using a sample drawn from U. An extension of the theory developed here to multinomial and cumulative logistic regression models is possible, but avoided for simplicity.

Poisson regression is often used when the dependent variable is restricted to positive values. This restriction can be, but does not have to 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/o_k.

Few additional assumptions about the distribution and variance structure of the ε_k are needed in the above vaguely-specified version of the model until the issue of estimating the variance of an estimator for $\boldsymbol{\beta}$ arises. That is a subject taken up in the next section.

Although apparently very general, there is a key restriction imposed by the standard model in equation (2.1); namely, that the expected value of the error term ε_k is zero no matter the value of $\mathbf{z}_k$. This assumption can fail. As a result, the standard model would not apply to the population. For example, suppose $y_k = z_k^2$ in the population. If one tries to fit the linear model $y_k = \alpha + \beta z_k + \varepsilon_k$ to this population, the standard model assumption $E(\varepsilon_k | \mathbf{z}_k) \neq 0$ for all realized $\mathbf{z}_k = (1\ z_k)^T$, $k \in U$, would fail.

A generalization of the standard model is the *extended model* under which $E(\varepsilon_k | \mathbf{z}_k) = 0$ in equation (2.2) is replaced by

$$E(\mathbf{z}_k \varepsilon_k) = \mathbf{0}. \tag{2.3}$$

In other words, ε_k has mean zero unconditionally (i.e., $E(\varepsilon_k) = 0$) and is uncorrelated with each of the components of $\mathbf{z}_k$. Unlike the standard model, the more general extended model rarely fails, so long as the first three central moments of the components of $\mathbf{z}_k$ are finite.

Suppose $y_k = z_k^2$ holds throughout the population, but we try to fit the population with the linear model $y_k = \alpha + \beta z_k + \varepsilon_k$ when $y_k = z_k^2$. The population analogue of ordinary least squares reveals $\beta = \text{Cov}(z_k^2, z_k)/ \text{Var}(z_k)$ and $\alpha = E(z_k^2) - \beta E(z_k)$. If the z_k were uniformly distributed on $U = [0, 1]$, then α would be -1/6 and β would be 1. Consequently, both $E(\varepsilon_k | z_k = 0) = 0 - (-1/6)$ and $E(\varepsilon_k | z_k = 1) = 1 - (5/6)$ would be positive, while $E(\varepsilon_k | z_k = 1/2) = 1/4 - 1/3$ would be negative. The $E(\varepsilon_k)$ and $E(z_k \varepsilon_k)$, by contrast, would both be 0.

Observe that the standard version of simple linear model through the origin, $y_k = \beta z_k + \varepsilon_k$, is not exactly of the form specified by equation (2.1) because it is missing an intercept. It similarly assumes $E(\varepsilon_k | z_k) = 0$. The extended version of this model assumes only $E(\varepsilon_k) = 0$.

2.1 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:

$$y_k = (q_k \boldsymbol{\delta}_k^T)\, \boldsymbol{\beta} + \varepsilon_k, \tag{2.4}$$

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 (2.4) 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 (2.4) 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 (2.4), the value β_g is the mean of g^{th} group, also called the *domain mean* of group g. In this monograph, we mostly treat population and domain means and their analogous ratios as parameters of a linear regression even though there are separate procedures in SUDAAN that can be used for their estimation.

Unlike the group-mean model, the group-ratio model does not fit our formal definition of a regression model in equation (2.1) 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 (e.g. when $q_k \equiv 1$, $z_{k1} + z_{k2} + \ldots + z_{kG} = 1$). Equation (2.2) is effectively replaced by $E(\varepsilon_k | \boldsymbol{\delta}_k) = 0$ for all realized $\boldsymbol{\delta}_k$, $k \in U$.

So long as the y_k are bounded, the group-mean and group-ratio models never fail.

2.2 The Weighted Estimating Equation

We will for the most part 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 usually 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 often 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, while 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 T_y. A sufficient condition for the probability limit of t, which we denote $p\lim(t_y)$, to be T_y as the population and expected 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

$$p\lim\left\{ M^{-1} \sum_{k \in U} \mathbf{z}_k \left[y_k - f(\mathbf{x}_k{}^T \boldsymbol{\beta}) \right] \right\} = p\lim\left\{ M^{-1} \sum_{k \in U} \mathbf{z}_k \varepsilon_k \right\} = \mathbf{0} \qquad (2.5)$$

under the extended model (where $E(\mathbf{z}_k \varepsilon_k) = \mathbf{0}$) 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 in the following section).

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

$$p\lim\left\{M^{-1}\sum_{k\in S} w_k \mathbf{z}_k\left[y_k - f(\mathbf{x}_k{}^T\boldsymbol{\beta})\right]\right\} = \mathbf{0} \tag{2.6}$$

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

$$p\lim \Psi_q = 0, \quad \text{where } \Psi_q = M^{-1}\left(\sum_{k\in S} w_k q_k - \sum_{k=1}^{M} q_k\right), \tag{2.7}$$

for $q_k = 1$, y_k, any component of $\mathbf{z}_k$, or any product of two of these. 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. More assumptions about the sample design will be made in the next section.

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

The analysis weights w_k may not exactly be 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 (see, for example, Kott 2006). Except in Section 3.1 on variance estimation via linearization, we treat the w_k as nearly equal to the inverse of the probability element k is jointly in 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 $k \notin S$. Until Section 6, we assume there is no item nonresponse.

The w_k are inserted into equation (2.6) in case $E(\varepsilon_k \mid w_k) \neq 0$, 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. 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, $E(\varepsilon_k \mid w_k) = 0$ 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 and Thompson 1974)

$$\sum_{k\in S} w_k \mathbf{z}_k\left[y_k - f(\mathbf{z}_k{}^T\mathbf{b})\right] = \mathbf{0} \tag{2.8}$$

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

$$\mathbf{b} - \boldsymbol{\beta} = \left[M^{-1}\sum_{k\in S} w_k f'(\theta_k)\mathbf{z}_k\mathbf{z}_k{}^T\right]^{-1} M^{-1}\sum_{k\in S} w_k \mathbf{z}_k \varepsilon_k, \tag{2.9}$$

for some θ_k between $\mathbf{z}_k{}^T\mathbf{b}$ and $\mathbf{z}_k{}^T\boldsymbol{\beta}$. This is a consequence of the mean-value theorem. An

additional mild condition we assume is that

$$\mathbf{A}_\theta = M^{-1}\sum_{k\in S} w_k f'(\theta_k)\mathbf{z}_k\mathbf{z}_k{}^T \tag{2.10}$$

and its probability limit, **A***, have finite components and are positive definite. When $M^{-1}\sum_S w_k\mathbf{z}_k\varepsilon_k$ converges to 0 in probability as the expected sample size grows arbitrarily large (see equation (2.5)), **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 maximum-likelihood (ML) estimating equation for the population under the independent and identically distributed (*iid*) linear regression model and under logistic regression with independently sampled population elements. Nevertheless, the solution to equation (2.8) is *not* ML when the weights vary or the ε_k within primary sampling units are correlated. Instead, the **b** solving equation (2.8) is referred to as a *pseudo-ML* estimator for $\boldsymbol{\beta}$ (Skinner 1989).

2.3 Pseudo-ML and Pfeffermann-Sverckkov Weight Adjustment

The pseudo-ML estimating equation in Binder is

$$\sum_{k\in S} w_k\left(\frac{f'(\mathbf{z}_k{}^T\mathbf{b})}{v_k}\right)\mathbf{z}_k\left[y_k - f(\mathbf{z}_k{}^T\mathbf{b})\right] = \mathbf{0}, \tag{2.11}$$

where $v_k = \mathrm{E}(\varepsilon_k^2|\mathbf{z}_k)$ is known (up to a scaling constant), and $\mathrm{E}(\varepsilon_k\varepsilon_j|\mathbf{z}_k, \mathbf{z}_j) = 0$ for $k \neq j$. For ordinary least squares (OLS) 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 for $f'(\mathbf{z}_k{}^T\boldsymbol{\beta}) = f(\mathbf{z}_k{}^T\boldsymbol{\beta})$. Thus for all three: $v_k \propto f'(\mathbf{z}_k{}^T\boldsymbol{\beta})$. This is not the case for (GLS) 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 for a in equation (2.11) in 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 $\sum_S\left(f'(\mathbf{z}_k{}^T\mathbf{b})/v_k\right)\mathbf{z}_k\left[y_k - f(\mathbf{z}_k{}^T\mathbf{b})\right] = \mathbf{0}$.

Pfeffermann and Sverchkov (1999) point out that when the standard model holds, 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 $k \neq i$, a more efficient estimator than the solution to equation (2.8) would factor each weight w_k in (2.11) 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 it can be replaced by $e_k^2 = [y_k - f(\mathbf{z}_k{}^T\mathbf{b})]^2$.

A possible way to generate $\omega(\mathbf{z}_k)$ is as the predicted value of an unweighted Poisson regression of $w_k e_k^2/f'(\mathbf{z}_k{}^T\mathbf{b})$ on the components of $\mathbf{z}_k = (z_{1k}, \ldots, z_{Pk})^T$ and, perhaps, functions of those components (e.g., $\log(z_{1k})$). Poisson regression is recommended because $w_k e_k^2/f'(\mathbf{z}_k{}^T\mathbf{b})$ is always positive. Recall that in Poisson regression it is $\log[w_k v_k/f'(\mathbf{z}_k{}^T\mathbf{b})]$ that is being modeled as a linear function of components or functions of components.

When the standard model holds in the population and $E(\varepsilon_k \varepsilon_i | \mathbf{z}_k, \mathbf{z}_i) = 0$ for $k \neq i$ can be assumed (or is close to holding) but the v_k are not known, one can try the following:

1) Fit the estimating equation in (2.8) and compute the $e_k = y_k - f(\mathbf{z}_k^T \mathbf{b})$.

2) When it is reasonable to assume $v_k \propto f'(\mathbf{z}_k^T \boldsymbol{\beta})$, as is always the case in a logistic regression, fit w_k on functions of components of $\mathbf{z}_k$ via an unweighted Poisson regression Call the fitted value $\omega(\mathbf{z}_k)$.

Otherwise,

2') fit $w_k e_k^2 / f'(\mathbf{z}_k^T \mathbf{b})$ on functions of components of $\mathbf{z}_k$ via an unweighted Poisson regression, and call the fitted value $\omega(\mathbf{z}_k)$.

3) Refit the estimating equation in (2.8) with each w_k replaced by the *Pfeffermann-Sverchkov-adjusted analysis weight*: $w_k^{PS} = w_k / \omega(\mathbf{z}_k)$.

When the fit in step 2 is good, these steps should return more efficient estimators for the components of $\boldsymbol{\beta}$ than fitting equation (2.8) and stopping. We will call this three-step process or any variant of it (e.g., one using linear rather than Poisson regression in step 2) *P-S weight adjustment.*

3. 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 arbitrarily large or it can be fixed. In the former situation, the number of PSUs in a stratum is fixed, while in the later that number 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 or not the number of strata should be treated a fixed in an asymptotic framework depends on the design. A design with, say, 60 strata containing two sampled PSUs in each is more reasonably treated in an asymptotic framework where number of strata grows large, while 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), N (n) 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 of PSU j of stratum h. We assume that in every population in the sequence of populations:

$$M_{hj} \leq B_M < \infty \text{ for all } hj. \tag{3.1}$$

3.1 When First-Stage Stratification is Ignorable

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 will instead make following two ignorability assumptions about the stratum identifiers:

A. $E(\varepsilon_k|\mathbf{z}_k,\mathbf{u}_k) = 0$ ($E(\mathbf{z}_k\varepsilon_k|\mathbf{u}_k) = \mathbf{0}$ for the extended model); that is to say, the first-stage stratification is ignorable in expectation.
B. $E(\varepsilon_k\varepsilon_j|\mathbf{z}_k,\mathbf{u}_k,\mathbf{z}_j,\mathbf{u}_j) = 0$ (($E(\mathbf{z}_k\varepsilon_k\ \mathbf{z}_j\varepsilon_j|\mathbf{u}_k,\ \mathbf{u}_j) = 0$ 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 not unreasonable 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 since $\mathbf{z}_k$ in equation (2.1) 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 (2.9)

$$\mathbf{b}-\boldsymbol{\beta} = \left[\sum_{k\in S} w_k f'(\theta_k)\mathbf{z}_k\mathbf{z}_k{}^T\right]^{-1}\sum_{k\in S} w_k\mathbf{z}_k\varepsilon_k, \tag{3.2}$$

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, $\mathbf{A}^*$, have finite components and are and positive definite. We are assuming that $w_k = 1/\pi_k$ in variance estimation under the extended model until Section 3.3. For the standard model, the analysis weights can be scaled by a function of $\mathbf{z}_k$.

From equation (3.2), we can see the bias of $\mathbf{b}$ is nearly zero. Consequently, a good estimator for its mean squared error is also a good estimator for its variance.

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

$$\mathbf{var}(\mathbf{b}) =$$

$$\mathbf{D}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{j=1}^{n_h}\left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k - \frac{1}{n_h}\sum_{a=1}^{n_h}\sum_{\kappa\in S_{ha}} w_\kappa\mathbf{z}_\kappa e_\kappa\right)\left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k - \frac{1}{n_h}\sum_{a=1}^{n_h}\sum_{\kappa\in S_{ha}} w_\kappa\mathbf{z}_\kappa e_\kappa\right)^T\mathbf{D} \tag{3.3}$$

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

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 (2.10)), 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 form of $(M\mathbf{A}_0)^{-1}$.

Our assumptions assure the near unbiasedness of the variance estimator in equation (3.3) (as n grows arbitrarily large) given a sampling design and population such that $p\lim(n\Psi_q^2)$ is bounded, where Ψ_q is defined in equation (2.7). They also assure the near unbiasedness of

$$\mathbf{var}_A(\mathbf{b}) = \mathbf{D}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k\right)\left(\sum_{k\in S_{hj}} w_k\mathbf{z}_k e_k\right)^T\mathbf{D}. \tag{3.4}$$

From a model-based viewpoint, the keys to both variance estimators are that the

$$\mathbf{E}_{hj}^{\varepsilon} = \sum_{k\in S_{hj}} w_k\mathbf{z}_k\varepsilon_k \tag{3.5}$$

on the right-hand side of equation (2.9) have mean **0** and are uncorrelated across PSUs and that **A*** is the probability limit of $M^{-1}\mathbf{D}^{-1}$. The use of robust sandwich-type variance estimates like equations (3.3) and (3.4) (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 $\mathbf{A}_0$.

Additional variations of the variance/mean-squared-error estimator in equation (3.3) 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.

3.2 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 $\mathbf{E}_{hj}^{\varepsilon}$ are uncorrelated and have a common mean within strata had the first-stage PSUs been selected *with* replacement (which would have allowed the same PSU to be selected more than once, each selection treated as independent with independent subsampling of elements). Equation 3.3 (but not (3.4)) 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 (3.3) 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 (3.3) 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 (3.3) 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 (3.3) fails

to capture. Here, we avoid that problem by assuming 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 (3.3) 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 $\mathbf{var}(\mathbf{b})$ (the estimated variances of the components of $\mathbf{b}$) will tend to be no higher than and expected values of the diagonals of $\mathbf{var}_A(\mathbf{b})$. They will tend to be lower when some of the stratum means are non-zero. That is to say, the diagonals of $\mathbf{var}_A(\mathbf{b})$, if anything, tend to overestimate the variances of the components of $\mathbf{b}$.

From the above expression we can see that practice of collapsing "similar" strata into variance strata for variance-estimation purposes (using equation (3.3) 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 (3.3) requires each n_h to be at least 2.

When every PSU is a design stratum is selected into the sample, these *certainty* PSUs become the variance strata for use in equation (3.3) and the units chosen from them in the next stage of sampling (e.g., housing units selected from area clusters) are variance PSUs.

3.3 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 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, S_k is 1 when k is in the respondent sample, 0 otherwise, and $\mathbf{g}$ (if it exists) satisfies the following calibration equation:

$$\sum_{k \in U} w_k \mathbf{c}_k = \sum_{k \in U} d_k S_k q(\mathbf{x}_k^{\ T} \mathbf{g}) \mathbf{c}_k = \mathbf{T}_\mathbf{c}, \tag{3.6}$$

where $\mathbf{c}_k$ is a vector of calibration variables with the same number of components as $\mathbf{x}_k$. In practice, the two are often identical. The population total of $\mathbf{c}_k$ – or a nearly unbiased estimate of that total – is known and denoted $\mathbf{T}_\mathbf{c}$.

When equation (3.6) is used create calibrated weights that account for unit nonresponse, the components of $\mathbf{T}_\mathbf{c}$ can be estimates based on the sample before unit nonresponse, that

is $\mathbf{T_c} = \sum_U d_k c_k$. The probability of (unit) response is assumed to have the form $1/q(\mathbf{x}_k{}^T\boldsymbol{\gamma})$, and the $\mathbf{g}$ 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 the probability an element responds when sampled, is Poisson, that is, independent across the elements of the population. Accordingly, the respondent sample can be treated as a stratified multistage sample.

When equation (3.6) is used to calibrate weights that account for coverage error, $1/q(\mathbf{x}_k{}^T\mathbf{g})$ estimates the expected number of time 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 assume duplication doesn't occur in the frame, and the number of times k is in the sampling frame (0 or 1) is independent across population elements, so that 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.

When the calibration factor is not used for selection modeling but to increase the efficiency of estimated means and totals $q(t)$, it is often set at $1 + t$ (linear calibration), exp(t) (raking), or $1/(1 + t)$ (pseudo-empirical likelihood) and $\boldsymbol{\gamma} = \mathbf{0}$. Linear calibration and raking are also often 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 truncated logistic 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 those used to justify equation (2.9) and the consistency of $\mathbf{b}$,

$$\mathbf{g} - \boldsymbol{\gamma} = \left[M^{-1} \sum_{k \in U} d_k S_k q'(\varphi_k) \mathbf{c}_k \mathbf{x}_k{}^T \right]^{-1} M^{-1} \left[\mathbf{T_c} - \sum_{k \in U} d_k S_k q(\mathbf{x}_k{}^T\boldsymbol{\gamma}) \mathbf{c}_k \right]$$

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

Often many of the components of $\mathbf{x}_k$ will also be component of $\mathbf{z}_k$. If they *all* were or if we replace the standard-model assumption in equation (2.2) by

$$\mathrm{E}(\varepsilon_k | \mathbf{z}_k, \mathbf{x}_k) = 0 \text{ for all realized } \mathbf{z}_k \text{ and } \mathbf{x}_k, k \in U, \tag{3.7}$$

then it is easy to see from

$$\begin{aligned}\mathbf{b}-\boldsymbol{\beta} &= \left[M^{-1}\sum_{k\in U} w_k f'(\theta_k)\mathbf{z}_k\mathbf{z}_k{}^T \right]^{-1} M^{-1}\sum_{k\in U} w_k\mathbf{z}_k\varepsilon_k \\ &\approx \left[M^{-1}\sum_{k\in U} d_k S_k q(\mathbf{x}_k{}^T\mathbf{g}) f'(\mathbf{z}_k{}^T\mathbf{b})\mathbf{z}_k\mathbf{z}_k{}^T \right]^{-1} M^{-1}\sum_{k\in U} d_k S_k q(\mathbf{x}_k{}^T\mathbf{g})\mathbf{z}_k\varepsilon_k\end{aligned} \tag{3.8}$$

that equation (3.3) 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 (3.7) collapses to that of the standard model in equation (2.2) when the component 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 (3.7) 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 $\mathbf{g}$ on the right-hand side of equation (3.8). It may be approximately equal to $\boldsymbol{\gamma}$, but the approximation is not close enough to be ignored.

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

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

Let ξ_k be the p^{th} component of $M^{-1}\mathbf{A}^{*-1}\mathbf{z}_k\varepsilon_k$, so that the error of the p^{th} component of $\mathbf{b}$ is

$$\begin{aligned}\sum_{k\in S} w_k\xi_k &= M^{-1}\sum_{k\in U} d_k\left\{c_k^T\boldsymbol{\delta}^* + S_k(\xi_k - c_k^T\boldsymbol{\delta}^*)q(\mathbf{x}_k^T\mathbf{g})\right\} \\ &\approx M^{-1}\sum_{k\in U} d_k\left\{c_k^T\boldsymbol{\delta}^* + S_k(\xi_k - c_k^T\boldsymbol{\delta}^*)q(\mathbf{x}_k^T\boldsymbol{\gamma}) + S_k(\xi_k - c_k^T\boldsymbol{\delta}^*)q'(\mathbf{x}_k^T\boldsymbol{\gamma})\mathbf{x}_k^T(\mathbf{g}-\boldsymbol{\gamma})\right\} \\ &\approx M^{-1}\sum_{k\in U} d_k\left\{c_k^T\boldsymbol{\delta}^* + S_k(\xi_k - c_k^T\boldsymbol{\delta}^*)q(\mathbf{x}_k^T\boldsymbol{\gamma})\right\},\end{aligned} \tag{3.9}$$

where

$$\boldsymbol{\delta}^* = p\lim\left\{\left(\sum_{j\in U} d_j S_j q'(\mathbf{x}_j^T\mathbf{g})\mathbf{x}_j\mathbf{c}_j^T\right)^{-1}\sum_{j\in U} d_j S_j q'(\mathbf{x}_j^T\mathbf{g})\mathbf{x}_j\xi_j\right\}. \tag{3.10}$$

Dropping the $M^{-1}\sum_U d_k\left\{S_k(\xi_k - c_k^T\boldsymbol{\delta}^*)q'(\mathbf{x}_k^T\boldsymbol{\gamma})\mathbf{x}_k^T(\mathbf{g}-\boldsymbol{\gamma})\right\}$ term above requires asymptotic theory. Both $\mathbf{x}_k{}^T(\mathbf{g}-\boldsymbol{\gamma})$ and $M^{-1}\sum_U d_k\left\{S_k(\xi_k - c_k^T\boldsymbol{\delta}^*)q'(\mathbf{x}_k^T\boldsymbol{\gamma})\right\}$ are $O_p(1/\sqrt{n})$ under mild conditions, so their product is $O_p(1/n)$, which small enough to be ignored.

Because the probability of response is Poisson, we can treat the sample as a stratified multistage design, with $\tilde{\xi}_k = M^{-1}\left\{c_k^T\boldsymbol{\delta}^* + S_k(\xi_k - c_k^T\boldsymbol{\delta}^*)q(\mathbf{x}_k^T\boldsymbol{\gamma})\right\}$ as the element values in an asymptotically equivalent expression for the error of the p^{th} coefficient of $\mathbf{b}$. The linearized version of the variance estimator for the coefficient replaces $\boldsymbol{\delta}^*$ in $\tilde{\xi}_k$ with the expression within the curly brackets of equation (3.10) and $\boldsymbol{\gamma}$ with $\mathbf{g}$.

With replication, we do an asymptotically equivalent thing, but without having the compute some of the complicated terms in equations (3.9) and (3.10) when the standard model fails. Instead, replicate versions of **g** are computed in replicate calibration equations (3.6). In the next section, we explore one such replication technique: the jackknife.

When calibrating to a constant $\mathbf{T_c}$ in equation (3.6) (or, more appropriately, both sides of equation (3.6) divided by M, where $M^{-1}\mathbf{T_c}$ remains constant as the population size grows), $\tilde{\xi}_k$ becomes $M^{-1}\left\{S_k(\xi_k - c_k^T\boldsymbol{\delta}^*)q(\mathbf{x}_k^T\boldsymbol{\gamma})\right\}$ because $\sum_U d_k c_k^T\boldsymbol{\delta}^*$ is replaced by a constant $\mathbf{T_c}\boldsymbol{\delta}^*$, which does not contribute to the variance. Moreover, some of the components of $\mathbf{T_c}$ can come from the full sample and some components be constants or provided from outside samples.

4. Jackknife Variance Estimation

Replication techniques provide alternative methods for estimating the variance of **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 jackknife.

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

$$\sum_{k\in S} w_k^{(hj)}\mathbf{z}_k\left[y_k - f(\mathbf{z}_k{}^T\mathbf{b}^{(hj)})\right] = \mathbf{0}, \tag{4.1}$$

where $w_k^{(hj)} = 0$ when $k \in S_{hj}$
$w_k^{(hj)} = [n_h/(n_h-1)]\, w_k$ when $k \in S_{h+}$ but $k \notin S_{hj}$
$w_k^{(hj)} = w_k$ otherwise.

This is identical to the estimator for $\boldsymbol{\beta}$ (in equation (2.7) 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 (2.9) and (2.10)

$$\mathbf{b}^{(hj)} - \boldsymbol{\beta} = \left[\mathbf{A}_\theta^{(hj)}\right]^{-1} M^{-1}\sum_{k\in S} w_k^{(hj)}\mathbf{z}_k\varepsilon_k, \tag{4.2}$$

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 A*, just like $\mathbf{A}_\theta$. Consequently,

$$\mathbf{b}^{(hj)} - \mathbf{b} \approx \left[\mathbf{A}^*\right]^{-1} M^{-1}\left(\sum_{k\in S} w_k^{(hj)}\mathbf{z}_k\varepsilon_k - \sum_{k\in S} w_k\mathbf{z}_k\varepsilon_k\right)$$

$$= \left[\mathbf{A}^*\right]^{-1} M^{-1} \left(\frac{1}{n_h - 1} \sum_{\substack{k \in S_{h+} \\ k \notin S_{hj}}} w_k \mathbf{z}_k \varepsilon_k - \sum_{k \in S_{hj}} w_k \mathbf{z}_k \varepsilon_k \right)$$

$$= \left[\mathbf{A}^*\right]^{-1} M^{-1} \frac{n_h}{n_h - 1} \left(\frac{1}{n_h} \sum_{k \in S_{h+}} w_k \mathbf{z}_k \varepsilon_k - \sum_{k \in S_{hj}} w_k \mathbf{z}_k \varepsilon_k \right),$$

and some more algebra reveals that the *delete-1 jackknife variance estimator* for **b**,

$$\mathbf{var}_{\mathrm{D1J}}(\mathbf{b}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\mathbf{b}^{(hj)} - \mathbf{b})(\mathbf{b}^{(hj)} - \mathbf{b})^T, \tag{4.3}$$

is nearly equal to the (Taylor-series) linearization variance estimator in equation (3.3).

There are two main differences between $\mathbf{var}(\mathbf{b})$ in equation (3.3) and $\mathbf{var}_{\mathrm{D1J}}(\mathbf{b})$ in equation (4.3). The former replaces the $\varepsilon_{\mathbf{k}}$ with e_k. This often causes $\mathbf{var}(\mathbf{b})$ to slightly underestimate the variances of the components of **b** when the number of sampled PSUs in not "arbitrarily large," that is, in actual 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 **b** when the number of sampled PSUs in 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

$$\mathbf{var}_{\mathrm{D1J\text{-}alt}}(\mathbf{b}) = \sum_{h=1}^{H} (\mathbf{b}^{(h1)} - \mathbf{b})(\mathbf{b}^{(h1)} - \mathbf{b})^T \ . \tag{4.4}$$

4.1 The Delete-a-Group Jackknife

Several computed 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 which is

$$\mathbf{var}_{\mathrm{GJ}}(\mathbf{b}) = \sum_{r=1}^{R} M_r (\mathbf{b}^{(r)} - \mathbf{b})(\mathbf{b}^{(r)} - \mathbf{b})^T, \tag{4.5}$$

where each $\mathbf{b}^{(r)}$ is computed with its own set of replicate weights. Observe that equations (4.3) and (4.4) have this generic form.

To run a *delete-a-group (DAG) jackknife,* one first sorts the variance PSUs by variance stratum and assign each variance PSU systematically to one of *R replicate groups*, which are *not* the same thing as 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 (4.4). 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:

$$\begin{aligned} w_k^{(r)} &= 0 && \text{when } k \in S^{hr} \\ &= w_k n_h / \left(n_h - n_{hr}\right) && \text{when } k \notin S^{hr}, \end{aligned}$$

which explains the name.

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

$$\begin{aligned} w_k^{(r)} &= w_k && \text{when } S^{hr} \text{ is empty} \\ &= w_k \left[1 - (n_h - 1)Z\right] && \text{when } k \in S^{hr} \\ &= w_k \left(1 + Z\right) && \text{when } S^{hr} \text{ is not empty, and } k \notin S^{hr}, \end{aligned}$$

where $Z = \sqrt{\dfrac{R}{R-1}\dfrac{1}{n_h(n_h-1)}}$.

The proof that DAG jackknife work can be found in Kott (2001). All replicate variance estimators having a form like equation (4.5) may exhibit a slight tendency to have an upward bias (which shrinks to zero as the number of sampled PSUs grows arbitrarily large) due to $\mathbf{b}^{(r)}$ 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 using calibrated weights, if some of the components of $\mathbf{T}_\mathbf{c}$ in equation (3.6) come from an outside, independently drawn samples, each with the same number of DAG jackknife replicate groups as S, and the rest are either constants or come from S before unit nonresponse, then a DAG jackknife variance estimator can capture both the variance contributed from the outside sample and from S (and from adjusted for unit nonresponse).

5. Some Tests

5.1 Tests for Choosing Weights

Suppose an analyst wants to determine whether **b** and **b'**, each computed with its own sets of weights, are estimating the same thing. For example, to test whether weights are ignorable in expectation, the analyst could compare **b** computed using inverse-selection-probability analysis weights with **b'** computed using equal weights. If the estimated coefficient vectors are not significantly different, then weights might be ignored. Similarly, **b** computed with analysis weights could be compared with **b'** computed using P-S adjusted weights. This would provide an indirect test of the standard model, since using the P-S adjusted weights produces a consistent estimator for $\boldsymbol{\beta}$ under the standard model but not necessarily when the standard model fails.

Under the null hypothesis that **b** and **b'** are estimating the same thing, $\chi_r^2 = (\mathbf{b}-\mathbf{b}')^T \left[\mathbf{var}(\mathbf{b}-\mathbf{b}')\right]^{-1}(\mathbf{b}-\mathbf{b}')$ is asymptotically chi-squared with r degrees of freedom, r is the dimension of $\mathbf{z}_k$, and **var**(.) is a variance estimator analogous to the one in equation (3.3). or (5.3). A popular probability-sampling-based test statistic for whether **b** and **b'** are estimating the same thing is

$$F_{r,d-r+1} = \left(\frac{d-k+1}{d}\right)\frac{(\mathbf{b}-\mathbf{b}')^T \left[\mathbf{var}(\mathbf{b}-\mathbf{b}')\right]^{-1}(\mathbf{b}-\mathbf{b}')}{r}, \tag{5.1}$$

where d is the *nominal* degrees of freedom, that is, $n - H$. The F test in equation (5.1) is called the *adjusted Wald F* test in SUDAAN 11 (RTI International, 2012; p. 217), which also offers a host of variations, of which the adjusted Wald F (Fellegi, 1980) and the Satterthwaite-adjusted F, based on Rao and Scott's (1981) Satterthwaite-adjusted chi-squared test, are the best (see Korn and Graubard, 1990). Note that this is an adjustment of the numerator degrees of freedom. It treats $n - H$ as the denominator degrees of freedom.

This test, proposed in Kott (1991) which owes much to the more assumption-dependent Hausmann (1978) test, is relatively easy to conduct using popular design-based software in the following manner. Two copies are made for each respondent in the data set. Both are assigned to the same PSU which accounts for their being strongly correlated in variance estimation. The first copy is assigned the weight used to compute **b** and the second the weight used to compute **b'**. The row vector of covariates $\mathbf{z}_k^T$ of the regression is replaced by ($\mathbf{z}_k^T$ $\mathbf{z}_k^T$) for the first copy and by ($\mathbf{z}_k^T$ $\mathbf{0}^T$) for the second. The regression coefficient is then

$$\tilde{\mathbf{b}} = \begin{pmatrix} \tilde{\mathbf{b}}^{(1)} \\ \tilde{\mathbf{b}}^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbf{b}' \\ \mathbf{b}-\mathbf{b}' \end{pmatrix},$$

Testing whether $\tilde{\mathbf{b}}^{(2)} = \mathbf{b}' - \mathbf{b}$ is significantly different from **0** is straight forward.

5.2 Another Test for the Standard Model

One way to test whether the standard model holds in a population is by testing whether using P-S adjusted weights yields significantly different regression-coefficient estimates from using inverse-selection-probability weights. Below we describe another test. It may prove useful for determining whether the standard logistic model holds in the population, which can be difficult to determine with a clustered sample (Graubard et al. 1998).

Compute **b** in equation (2.8) using inverse-selection-probability weights, calibrated weights, or P-S modified weights. Compute $f_k = f(\mathbf{z}_k^T\mathbf{b})$ which nearly equals $f(\mathbf{z}_k^T\boldsymbol{\beta})$. Apply design-sensitive software to the *linear* model: $E(y_k) = \alpha + \beta f_k + \gamma f_k^2$. If g, the estimator for γ, is significantly different from 0, then the standard model fails for the model in (2.2) because $E(\varepsilon_k|\mathbf{z}_k)$ is clearly not 0 (f_k being a function of $\mathbf{z}_k$, and the variance estimator being robust to the heteroskedasticity of the $y_k - \alpha - \beta f_k - \gamma f_k^2$). That g is not significantly different from 0 is necessary for the standard model to hold but not sufficient to establish that it holds. Observe that when the standard model holds a, the estimator for α, should also not be significantly different from 0. This suggests testing whether a and g are simultaneously not significantly different from 0.

6. Imputing missing item values with a regression model

In this section, we change focus from model fitting to prediction, in particular, to 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

$$t_y = \sum_{k \in S} w_k y_k R_k + \sum_{k \in S} w_k f(\mathbf{z}_k{}^T \mathbf{b})(1 - R_k), \tag{6.1}$$

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 the w_k in equation (6.1) by $w_k/\sum_S w_j$.

Suppose the standard regression model relating y_k to $f(\mathbf{z}_k{}^T\boldsymbol{\beta})$ in equations (2.1) and (2.2) holds, and the probability of item response for each unit respondent k is wholly a function of the components of $\mathbf{z}_k$. Then choosing for $\mathbf{b}$ in equation (6.1) a solution to the equation,

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

where $\Phi(\mathbf{z}_k)$ is any scalar function of $\mathbf{z}_k$ provides a nearly unbiased estimator for T_y in some sense.

When the standard model does not hold or the probably 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

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

or

$$\sum_{k \in S} w_k r_k \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T \mathbf{b}) \right] = \mathbf{0}, \tag{6.2}$$

where $w_k = \dfrac{I_k}{E(I_k \mid \cdot)}$ is the analysis weight,

$r_k = \dfrac{1 - E(R_k \mid y_k, \mathbf{z}_k{}^A)}{E(R_k \mid y_k, \mathbf{z}_k{}^A)} R_k$ 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 in the next subsection.

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

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

implies

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

and equation (6.1) provides a nearly unbiased estimated for T_y is some sense *whether or not the standard regression model holds*.

A common example of imputation with a regression model is imputation with the group ratio model in equation (2.4). When q_k varies across the k, $\mathbf{z}_k = q_k \boldsymbol{\delta}_k$ does not contain an intercept as we noted earlier. Nevertheless,

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

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

Observe that t_y in equation (6.1) is nearly unbiased in some sense when either,
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 $E(w_k|\cdot) = 1$), or,

2, the probabilities of item response are constant within each group (and $E(w_k|\cdot) = 1$).

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.

6.1 Assuming and Fitting an Item-Response Model

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

$$E(R_k \mid {}_i\mathbf{x}_k) = h({}_i\mathbf{x}_k{}^T {}_i\boldsymbol{\gamma}), \tag{6.3}$$

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 in Section 3.3.

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 (6.3) 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):

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

or its near equivalent:

$$\sum_{k\in U} w_k \mathbf{z}_k{}^0 \frac{R_k}{h({}_i\mathbf{x}_k{}^T {}_i\mathbf{g})} = \sum_{k\in U} w_k \mathbf{z}_k{}^0. \tag{6.4}$$

The size of $\mathbf{z}_k{}^0$ in practice 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 (6.4) (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 (6.3)) is to equal $\mathbf{z}_k{}^0$, 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 (6.2) 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 the assumed item response model fit using equation (6.4) is indeed the true item-response model.

If the standard regression model in equations (2.1) and (2.2) holds, the model errors (the ε_k in equation (2.1)) 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 (6.4) by $\omega_k = \omega(\mathbf{z}_k)$ the predicted value of a Poisson regression of $w_k r_k[y_k - f(\mathbf{z}_k{}^T\mathbf{b})]^2/f'(\mathbf{z}_k{}^T\mathbf{b})$ on appropriately chosen components of $\mathbf{z}_k$ and functions of those components. For double robustness to obtain, we may have to add ω_k to the components of $\mathbf{z}_k$ in equation (2.1), when it is not already a linear function of those components.

6.2 Nonignorable Item Nonresponse

When y_k is a component of ${}_i\mathbf{x}_k$ in the item-response model, that is, item nonresponse in nonignorable, things are a bit more complicated. Fitting equations (6.4) and then (6.2) to determine, in turn, r_k and then $\mathbf{b}$ will produce a nearly unbiased estimator for T_y when the item-response model in equation (6.3) is correctly specified.

If *both* the item-response model and the standard regression model in equations (2.1) and (2.2) are correctly specified, then $\mathbf{b}$ is a nearly unbiased estimator for $\boldsymbol{\beta}$. This can be softened a bit thanks to standard regression model assumption in equation (2.2). 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 $r_k = R_k[1 - \mathrm{E}(R_k|{}_i\mathbf{x}_k)]/\mathrm{E}(R_k|{}_i\mathbf{x}_k)]$ needs only to be correctly specified up to a function of $\mathbf{z}_k$. Consequently, if we fit $\mathrm{E}(R_k|\ {}_i\mathbf{x}_k)]$ with $1/[1 + \exp(y_k g_y + \mathbf{z}_k{}^T\mathbf{g}_z)]$, 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, after fitting $1/[1 + \exp(y_k g_y + \mathbf{z}_k{}^T\mathbf{g}_z)]$, g_y may only be close to a consistent estimator for γ_y, and so $\mathbf{b}$ would only be close to being nearly unbiased.

We can again, potentially increase the efficiency of $\mathbf{b}$ by dividing the $w_k r_k$ in equation (6.2) by $\omega_k = \omega(\mathbf{z}_k)$ the predicted value of a Poisson regression of $w_k r_k[y_k - f(\mathbf{z}_k{}^T\mathbf{b})]^2/f'(\mathbf{z}_k{}^T\mathbf{b})$ on appropriately chosen components of $\mathbf{z}_k$ and functions of those components. For double robustness to obtain, we can, as before, add ω_k to the components of $\mathbf{z}_k$ in equation (2.1), when it is not already a linear function of those components.

6.3 Variance Estimation

The delete-a-group 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 (6.1), 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 (6.2). An efficiency increasing weighting factor, $1/\omega_k$, if it exists, 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 (6.1) 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 $f(\mathbf{z}_k{}^T\mathbf{b}) > 1/(2m)$, 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 delete-a-group jackknife replicate, it is the size of $f(\mathbf{z}_k^T\mathbf{b}^{(r)})$ that is compared to 1/(2m), …, or (2m – 1)/(2m).

When $f(.)$ is linear or Poisson, add the residual $y_j - f(\mathbf{z}_{kj}^T\mathbf{b})$ from one of the item respondents to $f(\mathbf{z}_k^T\mathbf{b})$ when y_k is missing. To choose which item respondent's residual to use as a donor for k, first sort the item respondents in random order and selected 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; then assign the residuals of the m selected donors randomly to the m item nonrespondents. In every jackknife replicate, the sample donor residual is used for a particular item nonrespondent when needed to avoid overestimating the contribution to variance from adding residuals to the imputation.

7. Discussion

Complex surveys are usually designed to estimate population totals, means, and simple ratios of collected survey items. Sometimes, however, analysts desire 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 on 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 of 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 components: 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 misspecified. 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 $\mathbf{b}$ as we saw in Section 2.3. In addition, so 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 **b** when the standard regression model *holds even when the function used to estimate the unit-response probability is misspecified.*

Indeed, when the standard regression model holds, if the inverse of the probability of selection into the respondent sample is a function of the regression model's explanatory variables, then one need not weight the estimating equation at all in computing a nearly unbiased **b**. 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, the more explanatory variables in in $\mathbf{z}_k$, the less the need for analysis weights in the estimating equation. Similarly, the more components in $\mathbf{z}_k$, the more likely the standard model is to hold. Section 5 describes 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 with 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 provide protection 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 (6.4) can be used to fit an item-response model. A calibration equation (3.6) 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 misspecified response models is a function of the regression model's explanatory variables, then using the fitted response model to create the analysis or item-response weight will produce nearly unbiased estimates. Although this last condition is not likely to be satisfied in practice, it suggests that using a misspecified response model may remove some potential for bias resulting from nonignorable nonresponse at the unit or item level.

The delete-a-group jackknife provides a useful method for estimating variances of coefficient estimates in a regression model or items 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, a vector $\mathbf{x}_k$ such that $E(\varepsilon_k | \mathbf{z}_k, \mathbf{x}_k) = 0$ for all realized $\mathbf{z}_k$ and $\mathbf{x}_k$.

References

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

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

Felligi, I. (1980). Approximate test of independence and goodness of fit based on stratified multistage surveys. *Journal of the American Statistical Association*, 75, 261–268.

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

Godambe, V.P. and Thompson, M.E. (1974). Estimating equations in the presence of a nuisance parameter. *Annals of Statistics*, 2, 568–571.

Graubard, B.I. and Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17, 73–96.

Graubard, B.I., Korn, E.L., and Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 170–174.

Hausman, J. (1978). Specification tests in econometrics, *Econometrica* 46, 1251–1271.

Korn, E.L. and Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *American Statistician*, 44, 270–276.

Kott, P. (2018). A design-sensitive approach to fitting regression models with complex survey data. *Statistics Surveys*, *12*, 1–17.

Kott, P. (2007). Clarifying some issues in the regression analysis of survey data. *Survey Research Methods*, 1, 11–18.

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

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

Kott, P. (1991). What does performing linear regression on sample survey data mean? *Journal of Agricultural Economics Research*, 30–33.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. (2nd ed.), New York: Wiley.

Pfeffermann, D. and 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.

Rao, J. and Scott, A. (1981). The analysis of categorical data from complex surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221–230.

Research Triangle Institute (2012). *SUDAAN Language Manual*, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In Skinner, C.J., Holt, D. and Smith, T.M.F. eds. *Analysis of Complex Surveys*. Chichester: Wiley, 59–87.