A design-sensitive approach to fitting regression models with complex survey data

Phillip S. Kott

RTI International
6110 Executive Blvd.
Rockville, MD 20852 USA
e-mail: pkott@rti.org

Abstract: Fitting complex survey data to regression equations is explored under a design-sensitive model-based framework. A robust version of the standard model assumes that the expected value of the difference between the dependent variable and its model-based prediction is zero no matter what the values of the explanatory variables. The extended model assumes only that the difference is uncorrelated with the covariates. Little is assumed about the error structure of this difference under either model other than independence across primary sampling units. The standard model often fails in practice, but the extended model very rarely does. Under this framework some of the methods developed in the conventional design-based, pseudo-maximum-likelihood framework, such as fitting weighted estimating equations and sandwich mean-squared-error estimation, are retained but their interpretations change. Few of the ideas here are new to the refereed literature. The goal instead is to collect those ideas and put them into a unified conceptual framework.

Keywords and phrases: Pseudo-maximum likelihood, extended model, proportional-odds model, generalized cumulative logistic model, design-based.

Received June 2017.

1. Introduction

The standard "design-based" framework for fitting a regression model to survey data was introduced by Fuller (1975) for linear regression and by Binder (1983) more generally. This framework treats the finite population as a realization of independent trials from a conceptual population. A maximum likelihood estimator for regression-model parameters could, in principle, be estimated from the finite-population values. Given a complex sample drawn from the finite population, either that impractical-to-calculate finite-population estimate or its limit as the finite population grows infinitely large is treated as the target of estimation in the Fuller/Binder design-based framework

That is not what most analysts think they are estimating when they fit regression models. We will explore an alternative model-based framework for estimating regression models introduced in Kott (2007) that is *sensitive* to the

complex sampling design and to the possibility that the usual model assumptions may not hold in the population. Under this design-sensitive robust model-based framework some of the methods developed in the conventional design-based framework, such as fitting weighted estimating equations and sandwich mean-squared-error estimation, are retained but their interpretations change. Few of the ideas in this paper are new to the refereed literature. The goal here is to collect those ideas and put them into a unified conceptual framework.

Section 2 lays out the standard and extended model which underlie the design-sensitive approach to fitting vaguely-specified regression model taken here as opposed to the more conventional design-based or pseudo-maximum-likelihood (pseudo-ML) approach seen in Skinner (1989), Binder and Roberts (2003), Pfeffermann (1993; 2011), Lumley and Scott (2015; 2017), and elsewhere. Section 3 describes the weighted estimating equation used to estimate model parameters in a consistent manner under the design-sensitive framework.

The design-sensitive framework described here is model based for the most part. Probability-sampling ("design-based") principles are invoked in setting up an asymptotic framework to more deeply robustify the model. A parallel framework for robustifying the regression model can be found in the econometric literature (e.g., White, 1980; 1984). Probability-based inference is also invoked in Section 6.2 to provide some justification for a mean-squared-error estimator.

The design-sensitive robust model-based framework is like the randomization-assisted model-based approach to estimating population means and totals proposed in Kott (2005) as a contrast to the popular model-assisted (design-based) approach advocated in Särndal et al. (1992). Kott replaced the conventional term "design" with the more accurate "randomization" ("probability-sampling" would have been even better) because some aspects of the sampling design, like stratification, often depend explicitly or implicitly on a model. Here we use the modifier "design-sensitive" because our robust model-based framework is quite literally sensitive to how the sample has been designed. In what follows, we shorten "design-sensitive robust model-based framework" to "design-sensitive framework" for convenience.

For linear and logistic regression estimation, there is no difference in the estimator under the design-sensitive and pseudo-ML approaches. As Section 4 shows, that is not the case when estimating a cumulative logistic model, where both approaches lead to consistent, but not identical, parameter estimators.

Section 5 discusses the alternative weights introduced in Pfeffermann and Sverchkov (1999). These produce consistent and potentially more efficient estimators under the standard model.

Section 6 describes sandwich-like mean-squared-error estimation (from Fuller, 1975 and Binder, 1983) under the design-sensitive framework in the absence of calibration weight adjustments. In their presence, replication is advocated. The section also discusses the not-completely-resolved issue of how to handle the stratification of primary sampling units. Sections 7 and 8 discuss tests of whether the standard model holds, and Section 9 offers some concluding remarks.

2. The design-sensitive robust model-based approach

We start by defining the *standard model* in the following robust (i.e, distribution-free) manner:

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

where

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

y_k is the dependent random variable being modeled in the population U, while $\mathbf{x}_k$ is a vector of r explanatory variables (covariates), one of which is 1, and $f(.)$ is a monotonic function. Observe that

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

Few additional assumptions about the distribution and variance structure of the ε_k are needed in this vaguely-specified version of the model underpinning a regression analysis until the issue of estimating the mean-squared error of an estimator of $\boldsymbol{\beta}$ arises, as we shall see in Section 6.

Although apparently very general, there is a key restriction imposed by the standard model in equation (2.2); namely, that the expected value of the error term ε_k is zero no matter the value of $\mathbf{x}_k$. This assumption can fail and the standard model not be appropriate in the population being analyzed. For example, suppose $y_k = x_k^2$ in the population. Suppose one tries to fit the linear model $y_k = \alpha + \beta x_k + \varepsilon_k$ to this population. The standard model assumption $\mathrm{E}(\varepsilon_k | \mathbf{x}_k) \neq 0$ for all realized $\mathbf{x}_k$, $k \in U$, would fail.

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

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

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

Consider fitting the linear model $y_k = \alpha + \beta x_k + \varepsilon_k$ when $y_k = x_k^2$ in the population. The population analogue of ordinary least squares reveals $\beta = \mathrm{Cov}(x_k^2, x_k) / \mathrm{Var}(x_k)$ and $\alpha = \mathrm{E}(x_k^2) - \beta \mathrm{E}(x_k)$. If the x_k were uniformly distributed on $U = [0, 1]$, then α would be $-1/6$ and β would be 1. Consequently, both $\mathrm{E}(\varepsilon_k | x_k = 0) = 0 - (-1/6)$ and $\mathrm{E}(\varepsilon_k | x_k = 1) = 1 - (5/6)$ would be positive, while $\mathrm{E}(\varepsilon_k | x_k = 1/2) = 1/4 - 1/3$ would be negative. The $\mathrm{E}(\varepsilon_k)$ and $\mathrm{E}(\varepsilon_k x_k)$, by contrast, would both be 0.

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

3. The asymptotic framework and the weighted estimating equation

Although populations from which survey 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 survey data. If we assume there is an infinite sequence of nested populations growing arbitrarily large and that a sample is drawn from each using the same selection mechanism (which includes the self-selection of responding to the survey and the possibly-imperfect quasi-selection of individual population units into the sampling frame), where the samples are not necessarily nested but also grow arbitrarily large, then it is possible to take the probability limit of a statistic based on a sample as the expected sample size and population grow arbitrarily large (formally as we advance *ad infinitum* from one population in the sequence to the next). "Expected sample size" is because under many selected mechanisms the sample size is not fixed.

Suppose t is an estimator for T. A sufficient condition for the probability limit of t, denoted $p\lim(t)$, to be T is for the limit of the mean-squared error of t to converge to 0, in which case t is a consistent estimator for T.

Letting N denote the size of population U, then it is not difficult to see that

$$p\lim\left\{N^{-1}\sum_U \left[y_k - f(\mathbf{x}_k^T\boldsymbol{\beta})\right]\mathbf{x}_k\right\} = p\lim\left\{N^{-1}\sum_U \varepsilon_k\mathbf{x}_k\right\} = \mathbf{0} \qquad (3.1)$$

under the extended model (where $\mathrm{E}(\varepsilon_k\mathbf{x}_k) = \mathbf{0}$) given mild assumptions about the values of the components of $\mathbf{x}_k$ (e.g., they are bounded in number and each have finite moments) and variance structure of the ε_k (e.g., they are uncorrelated across primary sampling units, each of which consists of a fraction of the population that converges to 0 as the population grows arbitrarily large). Given a complex sample S with weights $\{w_k\}$, each nearly equal to the inverse of the corresponding element's selection probability (i.e., any difference tends to zero as the expected sample size grows arbitrarily large),

$$p\lim\left\{N^{-1}\sum_S w_k \left[y_k - f(\mathbf{x}_k^T\boldsymbol{\beta})\right]\mathbf{x}_k\right\} = \mathbf{0} \qquad (3.2)$$

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

$$p\lim \Psi_{\mathbf{z}} = 0, \quad \text{where } \Psi_{\mathbf{z}} = N^{-1}\left(\sum_{k\in S} w_k z_k - \sum_{k=1}^N z_k\right), \qquad (3.3)$$

for $z_k = 1$, y_k, any component of $\mathbf{x}_k$, or any product of two of these. Sufficient additional assumptions are that the z_k have finite moments and that the joint probability of drawing two primary sampling units in the first-stage of random sampling is bounded by the product of their individual first-stage selection probabilities.

The modifier "nearly" needs to be added to "equal" when weights are calibrated either to increase the statistical efficiency of the resulting estimators (as in Deville and Särndal, 1992) or to account for unit nonresponse or frame-coverage errors (the element is missing from the frame or is contained on the frame more than once) as in Kott (2006). The latter set of calibration adjustments is based on fits of selection models. Consequently, even assuming (as we do) that the selection models employed (for nonresponse and/or coverage adjustment) are correct, the calibrated weights are only asymptotically equal to the inverses of the selection probabilities. We return to the issues surrounding calibrated weights in Section 6.3.

The w_k are inserted into equation (3.2) in case $\mathrm{E}(\varepsilon_k|\mathbf{x}_k, w_k) \neq 0$, a situation in which the weights are said to be *nonignorable in expectation* (with respect to the model − a phrase that usually goes without saying). Full ignorability of the weights or, equivalently, of the selection probabilities in the sense of Little and Rubin (2002), obtains when the conditional $\varepsilon_k|\mathbf{x}_k$ are independent of the w_k. If the original random sample is selected with probability proportion to some component of $\mathbf{x}_k$, while the variance of $\varepsilon_k|\mathbf{x}_k$ is a function of that same component, then $\varepsilon_k|\mathbf{x}_k$ is clearly not independent of w_k, and the weights are not ignorable, but they could still be ignorable in expectation (i.e, $\mathrm{E}(\varepsilon_k|\mathbf{x}_k, w_k) = 0$ for every realized $\mathbf{x}_k$ and w_k, $k \in U$).

Whether the standard or extended model is assumed to hold in the population, solving for b in the weighted estimating equation (Godambe and Thompson, 1974)

$$\sum_{\mathrm{S}} w_k[y_k - f(\mathbf{x}_k^{\mathrm{T}}\mathbf{b})]\mathbf{x}_k = 0 \tag{3.4}$$

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

$$\mathbf{b} - \boldsymbol{\beta} = \left[N^{-1}\sum_{S} w_k f'(\theta_k)\mathbf{x}_k\mathbf{x}_k^T\right]^{-1} N^{-1}\sum_{S} w_k\mathbf{x}_k\varepsilon_k, \tag{3.5}$$

for some θ_k between $\mathbf{x}_k^{\mathrm{T}}\mathbf{b}$ and $\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}$. This is a consequence of the mean-value theorem. In addition to equation (3.3), the mild conditions assume that

$$A_\theta = N^{-1}\sum_{\mathrm{S}} w_k f'(\theta_k)x_k x_k^{\mathrm{T}} \tag{3.6}$$

and its probability limit are positive definite. Because $N^{-1}\sum_{S} w_k\mathbf{x}_k\varepsilon_k$ converges to 0 in probability as the expected sample size grows arbitrarily large (equation (3.1)), **b** is a consistent estimator for $\boldsymbol{\beta}$.

It is not hard to show that $\sum_{U}[y_k - f(\mathbf{x}_k^T\mathbf{b})]\mathbf{x}_k = \mathbf{0}$ 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 independent observations (i.e., sampled population elements). Nevertheless, the solution to equation (3.4) is *not* ML when the weights vary or the ε_k within primary sampling units are correlated. Instead, the **b** solving

(3.4) is referred to as *pseudo-ML* (Skinner, 1989). Strictly speaking, the full-population solution to $\sum_U [y_k - f(\mathbf{x}_k^T \mathbf{b})]\mathbf{x}_k = \mathbf{0}$ in linear or logistic regression need not be ML estimates under the design-based approach because the standard model may not hold in the population. Some advocates of model fitting using pseudo-ML, however, assume that the model does hold in the population (e.g., Pfeffermann, 1993; 2011).

4. The cumulative logistic model

More generally, the pseudo-ML estimating equation in Binder is

$$\sum_S w_k \frac{f'(\mathbf{x}_k^T \mathbf{b})}{v_k} \left[y_k - f(\mathbf{x}_k^T \mathbf{b}) \right] \mathbf{x}_k = \mathbf{0}, \tag{4.1}$$

where $v_k = \mathrm{E}(\varepsilon_k^2 | x_k)$ and $\mathrm{E}(\varepsilon_k \varepsilon_j | x_k, x_j) = 0$ for $k \neq j$. For logistic, Poisson, and ordinary least squares (OLS) linear regression, $f'(\mathbf{x}_k^T \boldsymbol{\beta})/v_k \propto 1$. This equality does not hold for general least squares (GLS) linear regression, however, where v_k varies across the elements of the population.

We will return to GLS linear regression in the next section. For now, we consider another example of when the pseudo-ML and design-sensitive estimators differ. The *cumulative logistic model* is a multinomial logistic regression model for ordered data, where there are L categories with a natural ordering (e.g., always, frequently, sometimes, never). Being in the first category is assumed to fit a logistic model. Being in either the first or second category is assumed to fit a different logistic model. Being in the first, second, or third category is assumed to fit yet another logistic model, and so forth.

The generalized cumulative logistic model is (splitting out the intercept from the rest of the covariates)

$$E(y_{\ell k} | \mathbf{x}_k) = \frac{\exp(\alpha_\ell + \mathbf{x}_k \boldsymbol{\beta}_\ell)}{1 + \exp(\alpha_\ell + \mathbf{x}_k \boldsymbol{\beta}_\ell)} \quad \text{for} \ell = 1, \ldots, L-1, \tag{4.2}$$

and $y_{\ell k} = 1$ if k is in one of the first l categories, 0 otherwise. When $\boldsymbol{\beta}_\ell = \boldsymbol{\beta}$ for all categories, but each category has its own intercept, the cumulative logistic model is also called a *proportional-odds model*.

Finding the $\mathbf{b}_\ell$ that satisfies the estimating equations:

$$\sum_S w_k \left[y_{\ell k} - f(\mathbf{x}_k^T \mathbf{b}_\ell) \right] \left[\begin{array}{c} 1 \\ \mathbf{x}_k \end{array} \right] = \mathbf{0} \quad \text{for} \ell = 1, \ldots, L-1 \tag{4.3}$$

can be used for the generalized cumulative logistic model or the proportional odds model. This is *not* the pseudo-ML estimating equation in the *surveylogistic* routine in SAS/STAT 14.1 (SAS Institute Inc., 2015), the *logistic* routine in SUDAAN 11 (Research Triangle Institute, 2012) or the *gologit2* routine in STATA (Williams, 2005).

When the standard model fails, that is, when

$$E\left[y_{\ell k} - \frac{\exp(\alpha_\ell + \mathbf{x}_k\boldsymbol{\beta}_\ell)}{1 + \exp(\alpha_\ell + \mathbf{x}_k\boldsymbol{\beta}_\ell)}\bigg|\mathbf{x}_k\right] \neq 0 \quad \text{for } \ell = 1, \ldots, L-1,$$

the solution for the $\mathbf{b}_\ell$ in equation (4.2) may not be estimating the same parameter as the pseudo-ML $\mathbf{b}_\ell^{PML}$. This is not a bad thing. Unlike the pseudo-ML solution, the solution to equation (4.3) has this reasonable property:

$$N^{-1}\sum_S w_k y_{\ell k} = N^{-1}\sum_S w_k f(\mathbf{x}_k^T\mathbf{b}_\ell) \quad \text{for } \ell = 1, \ldots, L-1.$$

This is a property retained at the asymptotic limit of $\mathbf{b}_\ell$ but not necessarily the asymptotic limit of $\mathbf{b}_\ell^{PML}$. That is to say,

$$\lim\left\{N^{-1}\sum_U w_k y_{\ell k}\right\} = \lim\left\{N^{-1}\sum_U f(\mathbf{x}_k^T\boldsymbol{\beta}_\ell)\right\} \quad \text{for } \ell = 1, \ldots, L-1,$$

where $\boldsymbol{\beta}_\ell$ is the estimand of $\mathbf{b}_\ell$. The equality need not hold when $\boldsymbol{\beta}_\ell$ is replaced by the estimand of $\mathbf{b}_\ell^{PML}$.

5. Modified weights under the standard model

5.1. Pfeffermann-Sverchkov (P-S) modified weights

Pfeffermann and Sverchkov (1999) showed that under the standard model one can replace the w_k with the *P-S modified weights* $w_{k\cdot g} = w_k g_k$ where g_k is a function of the components of $\mathbf{x}_k$, say $g(\mathbf{x}_k)$, computed to reduce as much as possible the variability of the $w_{k\cdot g}$ in the hopes of decreasing the variability of the linear regression-coefficient estimates under an *iid* model. Kott (2007) pointed out that the Pfeffermann-Sverchkov result is a simple repercussion of the assumption that $\mathrm{E}(\varepsilon_k|\mathbf{x}_k) = 0$ for all realized $\mathbf{x}_k$, $k \in U$. We can see that by replacing w_k in equations (3.2) and (3.4) by $w_k g(\mathbf{x}_k)$ and noting that $\mathrm{E}[g(\mathbf{x}_k)\varepsilon_k\mathbf{x}_k|\mathbf{x}_k] = \mathrm{E}(\varepsilon_k|\mathbf{x}_k) = 0$ for all realized $\mathbf{x}_k$, $k \in U$.

5.2. Listwise-deletion of observations with missing values

Interestingly, the Pfeffermann-Sverchkov result justifies the often reviled practice of deleting sampled observations with any missing values from a regression analysis (see, for example, Wilkinson et al. 1999). Under the standard model, *listwise deletion* leads to consistent coefficient estimates when the probability p_k that a sampled unit remains in the listwise-deleted sample is a function of the components of $\mathbf{x}_k$. (The probability an item value is missing is thus $1 - p_k$, which is also a function of the components of $\mathbf{x}_k$) Consequently, the true inverse-selection-probability weight is w_k/p_k, and a potential P-S modified weight is $w_k = w_k/p_k \times p(\mathbf{x}_k)$. Not only can item nonresponse be missing at random, explanatory-variable values can be missing not at random so long as their missingness does not depend on $y_k|\mathbf{x}_k$. Moreover, the functional form of $p(.)$ need not be known.

5.3. Generalized least squares

When $E(\varepsilon_k^2|x_k) = v_k$, $E(\varepsilon_k\varepsilon_j|x_k, x_j) = 0$ for $k \neq j$, and v_k is a function of x_k, then the pseudo-ML estimator for a linear regression estimator in equation (4.1) in consistent under the standard model (note that $f'(\mathbf{x}_k^T\mathbf{b}) = \mathbf{x}_k^T\mathbf{b}$). When the weights are ignorable, however, a more efficient estimator is the solution to

$$\sum_S \frac{\mathbf{x}_k^T\mathbf{b}}{v_k} \left[y_k - f(\mathbf{x}_k^T\mathbf{b})\right] \mathbf{x}_k = \mathbf{0}.$$

This suggests that when the weights are *not* ignorable a more efficient estimator than the solution to equation (4.1) would factor each weight w_k/v_k by $1/w(\mathbf{x}_k)$ where $w(\mathbf{x}_k)$ is a good approximation for w_k. A method for creating a reasonable form for $w(\mathbf{x}_k)$ is by setting $w(\mathbf{x}_k) = \exp(h(\mathbf{x}_k))$, where $h(x_k)$ is the fitted values of the ordinary-least-squares regression in the sample of $\log(w_k)$ on components of $\mathbf{x}_k = (x_{1k}, \ldots, x_{rk})^T$ and, perhaps, functions of those components (e.g., x_{1k}^2)

6. Mean-squared-error estimation

In this section, we restrict attention to stratified or single-stratum probability samples of primary sampling units (PSUs) of fixed size. Additional stages of probability samples can then 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 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.

Let h denote one of H strata, $\mathbf{u}_k$ the H-vector of stratum-inclusion identifiers for element k (i.e., if k is in stratum h, then u_{hk}, the hth component of $\mathbf{u}_{k,}$ is 1, while the other $H - 1$ components are 0), $N(n)$ 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

$$\lim_{N\to\infty} (M_{hj}/N) = 0 \quad \text{for all } hj. \tag{6.1}$$

6.1. When first-stage stratification is ignorable

Mean-squared-error estimation given a stratified multistage sample can be tricky unless a simplifying assumption is made. Usually, the assumption is that the PSUs were randomly selected *with replacement* within strata.

Instead, we assume for now that the ε_k are uncorrelated for elements from different PSUs, have bounded variances, and $E(\varepsilon_k|\mathbf{x}_k, \mathbf{u}_k) = 0$ ($E(\mathbf{x}_k\varepsilon_k|\mathbf{u}_k) = \mathbf{0}$ for the extended model) which is to say the first-stage stratification is ignorable in expectation. Although it is likely that strata were chosen such that the mean of the y_k differed across strata, it is nonetheless not unreasonable to assume that the $E(\varepsilon_k|\mathbf{x}_k)$ (or $E(\mathbf{x}_k\varepsilon_k)$) are unaffected by the first-stage stratum identifiers especially since $\mathbf{x}_k$ 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., $u_{hk}x_{jk}$).

To estimate the mean-squared error of the consistent estimator $\mathbf{b}$ (i.e., its probability limit tends to $\boldsymbol{\beta}$ as n grows arbitrarily large), one starts with

$$\mathbf{b} - \boldsymbol{\beta} = \left[N^{-1}\sum_S w_k f'(\theta_k)\mathbf{x}_k\mathbf{x}_k^T\right]^{-1} N^{-1}\sum_S w_k\mathbf{x}_k\varepsilon_k, \qquad (3.5)$$

for some θ_k between $\mathbf{x}_k^T\mathbf{b}$ and $\mathbf{x}_k^T\boldsymbol{\beta}$ and the additional assumption that $\mathbf{A}_\theta = N^{-1}\sum_S w_k f'(\theta_k)\mathbf{x}_k\mathbf{x}_k^T$ (equation (3.6)) and its probability limit are finite and positive definite. Let ω_k be the inverse of the selection probability of k. We will assume that $w_k = \omega_k$ until Subsection 6.3.

The design-based mean-squared-error estimator for $\mathbf{b}$ (from Binder, 1983) is

$$\mathbf{var}(\mathbf{b}) = N^{-2}\mathbf{D}\sum_{h=1}^H \frac{n_h}{n_h-1}\sum_{j=1}^{n_h}\left(\sum_{k\in S_{hj}} w_k\mathbf{x}_k e_k - \frac{1}{n_h}\sum_{\varphi=1}^{n_h}\sum_{\kappa\in S_{h\varphi}} w_\kappa\mathbf{x}_\kappa e_\kappa\right)$$
$$\left(\sum_{k\in S_{hj}} w_k\mathbf{x}_k e_k - \frac{1}{n_h}\sum_{\varphi=1}^{n_h}\sum_{\kappa\in S_{h\varphi}} w_\kappa\mathbf{x}_\kappa e_\kappa\right)^T \mathbf{D}, \qquad (6.2)$$

where $\mathbf{D} = \left[N^{-1}\sum_S w_k f'(\mathbf{x}_k^T\mathbf{b})\mathbf{x}_k\mathbf{x}_k^T\right]^{-1}$ estimates the definite probability limit of $\mathbf{A}_\theta$ in equation (3.6), and $e_k = y_k - f(\mathbf{x}_k^T\mathbf{b})$. Our assumptions assure the near unbiasedness of the variance estimator in equation (6.1) (as n grows arbitrarily lage) given a sampling design and the population such that $p\lim(n\Psi_{\mathbf{z}}^2)$ is bounded, where $\Psi_{\mathbf{z}}$ is defined in equation (3.2).

They also assure the near unbiasedness of

$$\mathbf{var}_A(\mathbf{b}) = \mathbf{D}\sum_{h=1}^H\sum_{j=1}^{n_h}\left(N^{-1}\sum_{k\in S_{hj}} w_k\mathbf{x}_k e_k\right)\left(N^{-1}\sum_{k\in S_{hj}} w_k\mathbf{x}_k e_k\right)^T\mathbf{D}. \quad (6.3)$$

Under the standard model (but not necessarily the extended model), the w_k in both equations (6.2) and (6.3) can be replaced by $w_{k\cdot g}$.

From a model-based viewpoint, the keys to both mean-squared-error estimators are that the $\mathbf{E}_{hj} = \sum_{k\in S_{hj}} w_k\mathbf{x}_k\varepsilon_k$ on the right-hand side of equation (3.5) have mean $\mathbf{0}$ and are uncorrelated across PSUs and that the probability limit of $\mathbf{D}$, like $\mathbf{A}_\theta$, is finite and positive definite. The use of robust sandwich-type

mean-squared-error estimates like equations (6.2) and (6.3) (the $\mathbf{D}$ being the bread of the sandwich) allows the variance matrices of the $\mathbf{E}_{hj}$ to be unspecified. The additional asymptotic assumptions allow $e_k = \varepsilon_k - \mathbf{x}_k^T(\mathbf{b} - \boldsymbol{\beta})$ to be used in place of ε_k within the $\mathbf{E}_{hj}$, and $\mathbf{D}$ to replace the probability limit of $\mathbf{A}_\theta$.

Additional variations of the mean-squared-error estimator in equation (6.2) 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 $\mathbf{x}_k$, the more reasonable the assumption that the ε_k are uncorrelated across elements (or another higher-stage of sampling like housing unit in a household-based sample of individuals) and the more reasonable the assumption that the first-stage stratification is ignorable.

6.2. When first-stage stratification is not ignorable

When the first-stage stratification is not ignorable and again $w_k = \omega$, it is tempting to follow design-based theory and argue that under probability-sampling theory the $\mathbf{E}_{hj}$ are independent and have a common mean within strata, justifying the use of the mean-squared-error estimator in equation (6.2) but not (6.3). This argument is only valid when the first-stage PSUs are indeed selected with replacement (which would allow the same PSU can be selected more than, with independent subsampling in each selection) or their selection can reasonably be approximated by that design.

A well-known problematic without-replacement sample is a systematic sample of elements drawn from a list ordered in cycles such that each possible sample severely over- or underestimates the estimand. Even when a systematic probability-proportional-size sample of PSUs is drawn from a randomly-ordered list, invoking large-sample or large- population properties (see Hartley and Rao, 1962) when the actual sample or population of PSUs in each stratum is not large is dubious.

Graubard and Korn (2002) point out that, even under with-replacement sampling of PSUs, equation (6.2) provides a nearly unbiased mean-squared-error estimator only when the relative sizes of the nonignorable strata are fixed as the population grows arbitrarily large. Otherwise, there is a component of the mean-squared error of $\mathbf{b}$ that equation (6.2) fails to capture: the random number of elements within each first-stage stratum when the number of strata is bounded (e.g., when the strata are the four US regions and the fraction of elements in each region is random).

6.3. An extension for calibrated weights

Let d_k denote the inverse of the probability that element k is chosen for the sample. The ratio ω_k/d_k is the product of calibration factors accounting for the probability of response which can occur at various levels (e.g., household and person for a survey of individuals) and the expected number of times the

element is in the sampling frame. When k is known not to be on the frame more than once, this value is the probability the k is on the frame at all.

To simplify the exposition, we will assume that there is a single calibration factor of the form $\omega_k/d_k = q(\mathbf{m}_k^T\boldsymbol{\gamma})$, where $q(t)$ is a monotonic function, such as $q(t) = 1 + \exp(t)$, $\mathbf{m}_k$ is a vector of variables with a finite number of components, $\boldsymbol{\gamma}$ is an unknown parameter consistently estimated by a $\mathbf{g}$ such that the following calibration equation holds:

$$\sum_S d_k q(\mathbf{m}_k^T\mathbf{g})\mathbf{c}_k = \mathbf{T_c}, \tag{6.4}$$

where $\mathbf{c}_k$ is a vector of calibration variables with the same number of components as – and often equal to – $\mathbf{m}_k$, and the population total of $\mathbf{c}_k$ or a nearly unbiased estimate of that total is known and denoted $\mathbf{T_c}$. When used to account for nonresponse, the components of $\mathbf{T_c}$ can be estimates based on the full sample before nonresponse.

When the calibration factor is used strictly 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 popular for coverage and nonresponse adjustment but $\boldsymbol{\gamma}$ is no longer $\mathbf{0}$. For nonresponse adjustment setting $q(t)$ at $1 + \exp(t) > 0$ assumes the probability of response is a logistic function of $\mathbf{m}_k$, while setting $q(t)$ at $(L + \exp(t))/(1 + \exp(t)/U)$ assumes response is a truncated logistic function with response probabilities between $1/U$ and $1/L$.

Under mild conditions, paralleling those used to justify equation (3.5) and the consistency of $\mathbf{b}$ and assuming the selection model for response/nonresponse or coverage error implied by $q(\mathbf{m}_k^T\boldsymbol{\gamma})$ is correct and that if $\mathbf{T_c}$ is a random variable then it is uncorrelated with whether element k is a respondent when sampled,

$$\mathbf{d} - \boldsymbol{\delta} = \left[N^{-1}\sum_S d_k q'(\varphi_k)\mathbf{c}_k\mathbf{m}_k^T\right]^{-1} N^{-1}\left[\mathbf{T_c} - \sum_S d_k q(\mathbf{m}_k^T\boldsymbol{\gamma})\mathbf{c}_k\right]$$

for some φ_k between $\mathbf{m}_k^T\mathbf{g}$ and $\mathbf{m}_k^T\boldsymbol{\gamma}$, and so $\mathbf{d}$ is a consistent estimator for $\boldsymbol{\delta}$. Just as in the near unbiasedness of $\mathbf{b}$, the differences between $1/q(\mathbf{m}_k^T\boldsymbol{\gamma})$ and the value it estimates (i.e., 1 if element k is a respondent, 0 if not; the number of times element k of the population is in the sampling frame) can be correlated within PSUs, although it is more common to assume that the differences are uncorrelated across all elements.

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

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

then it is easy to see from

$$\mathbf{b} - \boldsymbol{\beta} = \left[N^{-1}\sum_S w_k f'(\theta_k)\mathbf{x}_k\mathbf{x}_k^T\right]^{-1} N^{-1}\sum_S w_k\mathbf{x}_k\varepsilon_k$$

$$= \left[N^{-1} \sum_S w_k f'(\theta_k) \mathbf{x}_k \mathbf{x}_k^T \right]^{-1} N^{-1} \sum_S \omega_k q(\mathbf{m}_k^T \mathbf{g}) \mathbf{x}_k \varepsilon_k$$

that equation (6.3) (or (6.2)) can be used to estimate the mean-squared error of $\mathbf{b}$ given $\mathbf{T_c}$ when under the standard model. The conditioning is needed when $\mathbf{T_c}$ itself is an estimator.

More generally, we would be better served using a replication method (i.e., one that replicates the calibration factors so that equation (6.4) holds for each replicate) to estimate the (conditional) mean-squared error of $\mathbf{b}$ in a way that accounts for the mean-squared error contribution from $\mathbf{d}$ as well as the ε_k (and, perhaps, $\mathbf{T_c}$ as well). Although the mean-squared error with both these contributions can be linearized as well, linearization becomes increasingly cumbersome as the number of calibration factors increase. So too does replication because there is a calibration equation for each factor, but once replicate weights are determined they can be used for estimating any number of regression models. Linearization does not share this convenient property.

6.4. Degrees of freedom

When fitting a regression model to survey data, design-based practice often treats the diagonals of the mean-squared-error estimator in equation (6.3) as if they had a chi-squared distribution with $n - H$ degrees of freedom (Lohr, 2010, p. 438). There is no justification for this under probability sampling theory, which relies entirely on the asymptotic normality of $\mathbf{b}$. This questionable practice clearly comes from $\mathbf{var}(\mathbf{b})$ in equation (6.3) looking a bit like the multiple of a chi-squared statistic with $n - H$ degrees of freedom.

In fact, if the $\mathbf{E}_{hj}$ were all independent and identically distributed multidimensional normal random variables, then the diagonals of $\mathbf{var}(\mathbf{b})$ would indeed be close to a multiple of a chi-squared statistic with $n - H$ degrees of freedom. Unfortunately, the $\mathbf{E}_{hj}$ in practice are not likely to be normally distributed, and even if they are close enough to being normal for them to be treated as such, they rarely have the same variances.

A model-based approach in Kott (1994) assumes that the first-stage stratification is ignorable, and the ε_k (as opposed to the $\mathbf{E}_{hj}$) are normally distributed, have mean zero and a common variance, and are uncorrelated. The approximate relative mean-squared error of a diagonal of $\mathbf{var}(\mathbf{b}) \approx \mathbf{var}(N^{-1} \sum \sum \mathbf{DE}_{hj})$, call it rv, can be calculated under those assumptions. Using the Satterthwaite approximation, the *effective* degrees of freedom for the corresponding component of $\mathbf{b}$ would then be $2/rv$ and could vary across coefficients. It is

$$F = \frac{\left(\sum_{h=1}^H \sum_{j=1}^{n_h} \sum_{k \in s_{hj}} w_k^2 \right)^2}{\left(\sum_{h=1}^H \sum_{j=1}^{n_h} \sum_{k \in s_{hj}} w_k^4 + \sum_{h=1}^H \frac{1}{(n_h-1)^2} \sum_{j,g=1, j \neq g}^{n_h} \left[\sum_{k \in s_{hj}} w_k^2 \right] \left[\sum_{k \in s_{hg}} w_k^2 \right] \right)}.$$

Although this procedure is itself more than a little questionable, when it is employed in the generation of t statistics the result will likely produce better

coverage intervals than conventional design-based practice. Better yet would be to compute alternative measures of the effective degrees of freedom under different assumptions about the variance structure of the ε_k within a sensitivity analysis. Valliant and Rust (2010) address the degrees-of-freedom problem in design-based regression analyses.

7. Tests for choosing weights

Suppose an analyst wants to determine whether $\mathbf{b}$ and $\mathbf{b}'$, 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 $\mathbf{b}$ computed using inverse-selection-probability weights with $\mathbf{b}'$ computed using equal weights. If the vectors are not significantly different, then weights might be ignored. Similarly, $\mathbf{b}$ could be compared with a different $\mathbf{b}'$ computed using P-S modified weights. This would provide an indirect test of the standard model, since using the P-S modified weights produces a consistent estimator for $\boldsymbol{\beta}$ under the standard model but not more generally.

Under the null hypothesis that $\mathbf{b}$ and $\mathbf{b}'$ are estimating the same thing, $\chi_r^2 = (\mathbf{b} - \mathbf{b}')^T [\mathbf{var}(\mathbf{b} - \mathbf{b}')]^{-1} (\mathbf{b} - \mathbf{b}')$ is asymptotically chi-squared with r degrees of freedom, r is the dimension of $\mathbf{x}_k$, and $\mathbf{var}(.)$ is a mean-squared-error estimator analogous to the one in either equation (6.2) or (6.3). A popular design-based test statistic for whether $\mathbf{b}$ and $\mathbf{b}'$ are estimating the same thing is

$$F_{r,d-r+1} = \left(\frac{d - k + 1}{d} \right) \frac{(\mathbf{b} - \mathbf{b}')^T [\mathbf{var}(\mathbf{b} - \mathbf{b}')]^{-1} (\mathbf{b} - \mathbf{b}')}{r}, \qquad (7.1)$$

where d is the *nominal* degrees of freedom, that is, $n - H$. The F test in equation (7.1) 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 (Felligi, 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).

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

$$\mathbf{d} = \left(\begin{array}{c} \mathbf{b}' \\ \mathbf{b} - \mathbf{b}' \end{array} \right),$$

and testing whether $\mathbf{b}' - \mathbf{b}$ is significantly different from $\mathbf{0}$ becomes a straightforward regression exercise using design-based software, such as SUDAAN 11 or the survey procedures in SAS/STAT 14.1 (SAS Institute Inc., 2015).

The design-sensitive model-based approach allows each component of $\mathbf{d}$ to have its own model-based effective degrees of freedom in a t test and then uses a conservative Bonferroni adjustment to test whether the components in the bottom half of $\mathbf{d}$ are significantly different from 0 (i.e., the smallest p value among the components is compared to α/r when testing for significance at the α level). Using a Bonferroni-adjusted t test in place of an F test when analyzing a regression with complex survey data was previously advocated by Korn and Graubard (1990). Holm (1979) proposed a method for determining which components of $\mathbf{b}' - \mathbf{b}$ are significantly different from zero. Using the Holm-Bonferroni procedure, the component with the vth lowest p value is deemed statistically significant at the α level when the $v - 1^{\text{th}}$ lowest p value has been deemed significant and the vth lowest p value is less than $\alpha/(1 + r - v)$.

8. Another test for the standard model

Determining whether using P-S modified weights yields significantly different regression-coefficient estimates from using inverse-selection-probability weights is one way of testing whether the standard model holds. Here is another test that may prove useful for determining whether the standard logistic model holds in the population, which can be difficult with a clustered sample (Graubard et al., 1997).

Estimate $\mathbf{b}$ in equation (2.1) using inverse-selection-probability weights, calibrated weights, or P-S modified weights. Compute $f_k = f(\mathbf{x}_k^T \mathbf{b})$ which nearly equals $f(\mathbf{x}_k^T \boldsymbol{\beta})$. Apply "design-based" software to the *linear* model: $\mathrm{E}(y_k) = \alpha + \beta f_k + \gamma f_k^2$. If g, the estimator for γ, is significantly different from 0, then the standard model fails for the model in equation (2.1) because $\mathrm{E}(\varepsilon_k|\mathbf{x}_k)$ is clearly not 0 (f_k being a function of $\mathbf{x}_k$, and the "design-based" mean-squared-error 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.

9. Concluding remarks

The goal of this paper has been to show that some of the techniques in conventional design-based practice can be justified in the design-sensitive framework for estimating the vaguely-specified regression models defined herein. Nevertheless, although inserting weights into an estimating equation is often justified, it is not always necessary, depending on what assumptions are made. In addition, the sandwich-type mean-squared-error estimator used in design-based practice (equation (6.3)) does not fully account for first-stage stratification when first-stage stratification is not ignorable in expectation. When it is ignorable, a simpler mean-squared-error estimator (equation (6.2)) can be used that is likely

more stable (i.e., it diagonals have less relative mean-squared error). Other, even more stable, mean-squared-error estimators can be constructed if one is willing to assume that element errors are not correlated across smaller levels of clustering than PSUs (e.g., across households but not within households).

In practice the standard and extended models described here rarely produce estimators different from the popular pseudo-ML methodology. An exception to this is the cumulative logistic model. Ironically, it is a simple matter to employ SAS/STAT or SUDAAN to estimate a generalized cumulative logistic model using the methodology discussed here even though the analogous pseudo-ML estimator cannot be computed with SUDAAN. To do so one treats the L-1 equations involving the same element as if they different elements from the same PSU and runs a (binary) logistic regression, relying on the sandwich-like design-based mean-squared-error estimators to handle the correlation of the equations. Testing the "parallel lines" assumption of the proportional-odds model that all the $\boldsymbol{\beta}_\ell = \boldsymbol{\beta}$ in equation (4.1) is straightforward.

One interesting repercussion of assuming the robust standard model is that listwise deletion turns out to be an appropriate technique for regression analysis so long as the probability an element is deleted from the analysis does not depend on the value of the dependent variable given the independent variables.

The problem with making assumptions is that they can be wrong. Survey statisticians have, for the most part, accepted a design-based framework that effectively focuses on robustness by relying on as few model assumptions as possible. That framework is not particularly helpful when the goal is to fit a regression model. Moreover, it can be misleading when survey statisticians graft a notion like degrees of freedom onto what is an asymptotic theory.

The paper has reviewed statistical tests for determining whether inverse-selection-probability weights are ignorable in expectation when fitting a regression model and, if so, whether the standard model nonetheless holds, allowing the use of P-S modified weights. The practicality of these tests, which may not have much power, will need to be determined by further research.

Design-based practice has been to fear that any such test will incorrectly fail to see that the weights are not ignorable or that the standard model fails. In fact, the standard model, like all models, is almost never be completely true. In the same vein, inverse-selection-probability weights are rarely entirely ignorable. Still, the standard model may be useful, and the efficiency gains from ignoring the weights may overwhelm the resulting bias. We need better tools for making such determinations. The design-sensitive model-based approach may be the key to developing those tools.

References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279–292. MR0731144

Binder, D.A. and Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters, in R.L. Chambers and C.J. Skinner,

eds., *Analysis of Survey Data*, John Wiley & Sons, New York, Chapter 3. MR1978842

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382. MR1173804

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.

Fuller, W.A. (2002). Regression estimation for survey samples (with discussion). *Survey Methodology*, 28(1), 5–23.

Godambe, V.P. and Thompson, M.E. (1974). Estimating equations in the presence of a nuisance parameter. *Annals of Statistics*, 2, 568–571. MR0356335

Graubard, B.I. and Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17, 73–96. MR1910075

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.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling With Unequal Probabilities and Without Replacement. *Annals of Mathematical Statistics*, 33, 350–374. MR0143312

Hausman, J. (1978). Specification tests in econometrics, *Econometrica* 46, 1251–1271. MR0513692

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70. MR0538597

Korn, E. and Graubard, B. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from Survey data. *Survey Methodoloegy*, 193–201.

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

Kott, P.S. (1994). Hypothesis testing of linear regression coefficients with survey data. *Survey Methodology*, 20, 159–164.

Kott, P. (2005). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129, 263–277. MR2126850

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

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

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.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. (2nd ed.), New York: Wiley. MR1925014

Lohr, S. (2010). Sampling: Design and Analysis. (2nd ed.), Boston: Brooks/Cole. MR3057878

Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3, 1–18.

Lumley, T. and Scott, A. (2017). Fitting regression models to survey data. *Statistical Science*, 32, 265–278. MR3648959

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, 37(2), 115–136.

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

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. MR1720710

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. MR0624328

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

Särndal, C.-E., Swensson, B., and Wretmann, J. (1992). *Model assisted survey sampling*. New York: Springer. MR1140409

SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide*. Cary, NC: SAS Institute Inc.

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. MR1049386

Valliant, R. and Rust, K. (2010). Degrees of freedom approximations and rules-of-thumb. *Journal of Official Statistics*, 26, 585–602.

White, H. (1984). *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Review*, 21, 149–170. MR0572464

Wilkinson, L. and the Task Force on Statistical Inference, Board of Scientific Affairs, American Psychological Association (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 8, 594–604.

Williams, R. (2005). *Gologit2: A Program for Generalized Logistic Regression/Partial Proportional Odds Models for Ordinal Variables*. Retrieved January 3, 2016 (http://www.nd.edu/~rwilliam/stata/gologit2.pdf).