

A Design-Sensitive Approach to Fitting Regression Models With Complex Survey Data

Phillip S. Kott
RTI International
Rockville, MD

Introduction: What Does Fitting a Regression Model with Survey Data Mean?

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 regression estimator could, in principle, be estimated from the finite-population values. In the design-based framework either that uncalculated finite-population estimate or its limit as the finite population grows infinitely large is treated as the target of estimation given a complex sample drawn from the finite population.

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 framework some of the methods developed in the design-based framework, such as fitting weighted estimating equations and sandwich variance estimation, are retained but their interpretations change. Only a few of the ideas in this paper are new. The goal here is to collect those ideas and put them into a conceptual framework.

A Design-Sensitive Approach

We start by defining the *standard model* in the following manner:

$$y_k = f(\mathbf{x}_k^T \boldsymbol{\beta}) + \varepsilon_k, \quad \text{where } E(\varepsilon_k | \mathbf{x}_k) = 0, \quad (1)$$

where y_k is the dependent variable being modeled, while \mathbf{x}_k is a vector of variables, one of which is 1. Observe that

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

There are few additional assumptions about the distribution and variance structure of the ε_k in this robust version of the model underpinning a regression analysis until the issue of estimating the variance of an estimator of $\boldsymbol{\beta}$ arises.

Although apparently very general, there is key restriction imposed by the standard model in equation (1): $E(\varepsilon_k) = 0$ 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. The linear model $y_k = \alpha + \beta x_k + \varepsilon_k$ when fit to the population fails as a standard model because $E(\varepsilon_k | \mathbf{x}_k) \neq 0$.

A further generalized is the *extended model* under which $E(\varepsilon_k | \mathbf{x}_k) = 0$ in equation (1) is replaced by $E(\mathbf{x}_k \varepsilon_k) = \mathbf{0}$. That is to say, ε_k has mean zero unconditionally (i.e., $E(\varepsilon_k) = 0$) and is uncorrelated with each of the nonconstant components of \mathbf{x}_k . Unlike the standard model, the more general extended model rarely fails. Indeed, in the above example, $\beta = \text{Cov}(x_k^2, x_k) / \text{Var}(x_k)$ and $\alpha = E(x_k^2) - \beta E(x_k)$ so long as x_k the first three central moments of x_k are finite.

The standard version of simple linear model without an intercept, $y_k = \beta x_k + \varepsilon_k$, is not of the form specified by equation (1). It similarly assumes $E(\varepsilon_k | \mathbf{x}_k) = 0$. The extended version of the model assumes only $E(\varepsilon_k) = 0$.

The Weighted Estimating Equation

With an independent identically distributed (*iid*) population U of N elements, it is easy to see that

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

under the extended model. Given a complex sample S with weights $\{w_k\}$, each (nearly) equal to the inverse of the corresponding element's selection probability,

$$p \lim \left\{ N^{-1} \sum_S w_k \left[y_k - f(\mathbf{x}_k^T \boldsymbol{\beta}) \right] \mathbf{x}_k \right\} = \mathbf{0} \quad (2)$$

under mild conditions on the sampling design. The parenthetical “nearly” needs to be added when the weights include adjustments for unit nonresponse or coverage errors in the frame which the analysts assumes have been accounted for in an asymptotically unbiased manner. Calibration weight adjustments for statistical efficiency are another reason to add “nearly.”:

The w_k are inserting into equation (2) in case $E(\varepsilon_k | \mathbf{x}_k, w_k) \neq 0$ a situation in which the weights are said to be *nonignorable in expectation*. Full ignorability of the weights obtains when $\varepsilon_k | \mathbf{x}_k$ is independent of w_k .

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)

$$\sum_S w_k [y_k - f(\mathbf{x}_k^T \mathbf{b})] \mathbf{x}_k = \mathbf{0} \quad (3)$$

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

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

where θ_k is between $\mathbf{x}_k^T \mathbf{b}$ and $\mathbf{x}_k^T \boldsymbol{\beta}$ thanks to the mean-value theorem. The mild conditions include that

$$\mathbf{A}_0 = N^{-1} \sum_S w_k f'(\theta_k) \mathbf{x}_k \mathbf{x}_k^T$$

and its limit are positive definite, while $N^{-1} \sum_S w_k \mathbf{x}_k \varepsilon_k$ converges to 0 as the sample grows arbitrarily large.

It is not hard to show that $\sum_U [y_k - f(\mathbf{x}_k^T \boldsymbol{\beta})] \mathbf{x}_k = \mathbf{0}$ is the maximum-likelihood (ML) estimating equation of the population under the *iid* linear regression model and under logistic regression with independent observations (i.e., sampled elements). Nevertheless, the solution to equation (3) is *not* ML given only sample values when the weights vary or the ε_k within primary sampling units are correlated. Instead, it is referred to as *pseudo-ML* (Skinner 1989).

The Cumulative Logistic Model

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

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

For logistic, Poisson, and ordinary least squares (OLS) linear regression, $f'(\mathbf{x}_k^T \boldsymbol{\beta})/v_k = 1$. This equality may not hold for general least squares (GLS) linear regression, however even when the elements are uncorrelated. For uncorrelated GLS and known or speculated v_k up to a constant, one could choose g_k in $w_k g_k = w_k v_k$ to limit the variability of $w_k v_k g_k$ under the standard model.

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 logistic model. Being in the first, second, or third category is assumed to fit a logistic model, and so forth

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

$$E(y_{\ell k} | \mathbf{x}_k) = \frac{\exp(\alpha_{\ell} + \mathbf{x}_k \boldsymbol{\beta}_{\ell})}{1 + \exp(\alpha_{\ell} + \mathbf{x}_k \boldsymbol{\beta}_{\ell})} \quad \ell = 1, \dots, L-1, \quad (4)$$

and $y_k = 1$ 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}_{ℓ} that satisfies the estimating equation:

$$\sum_S w_k \left[y_{\ell k} - f(\mathbf{x}_k^T \mathbf{b}_{\ell}) \right] \begin{bmatrix} 1 \\ \mathbf{x}_k \end{bmatrix} = \mathbf{0} \quad \text{for } \ell = 1, \dots, L-1 \quad (5)$$

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). Only the last goes beyond the beyond the proportional-odds model.

When the standard model fails, that is, when

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

the solution for the \mathbf{b}_{ℓ} in equation (4) may not be estimating the same parameter as the pseudo-ML \mathbf{b}_{ℓ}^{PML} . This is not a bad thing. Unlike the pseudo-ML solution, the solution to equation (4) has this reasonable property:

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

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

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

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

Modified Weights Under the Standard Model

Pfeffermann and Sverchkov (1999) showed that under the standard model one can replace the w_k with the modified weights $w_{k-g} = w_k g_k$ where g_k is a function of the components of \mathbf{x}_k computed to reduce as much as possible the variability of the w_{k-g} in the hopes of decreasing the variability of the linear regression-coefficient estimates under an *iid* model. Kott (2007) pointed out that Pfeffermann's and Sverchkov's result is a simple repercussion of the assumption that $E(\epsilon_k | \mathbf{x}_k) = 0$. We can see that by replacing w_k in equations (2) and (3) by w_{k-g} .

Interestingly, Pfeffermann's and Sverchkov's result also 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 if the probability that a sampled unit remains in the listwise-deleted sample is a function of the components of \mathbf{x}_k , say p_k (and the probability an item value is missing is thus $1 - p_k$). As a result, the true inverse-selection-probability weight is w_k/p_k , and a potential modified weight is $w_k = w_k/p_k \times p_k$. Not only can item nonresponse be missing at random, independent-variable values can be missing not at random so long as their missingness does not depend on $y_k | \mathbf{x}_k$. Moreover, not even the function form of p_k need not be known.

Observe that the estimation of a mean or a domain mean can be put in the form of a linear-regression model (and often a logistic-regression model). The standard model can never fail in this case nor can it fail for the *group-mean model*, where the population is divided into groups and each group has its own mean. Nevertheless, the modified weights can't be used to reduce variability because g_k will be constant within the domain/group. Moreover, the weights need not be ignorable in expectation.

Variance Estimation 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 after primary sampling units (PSUs) had been separated into mutually exclusive strata, the PSUs were randomly selected *with replacement* within strata before elements were selected independently within sampled PSU using some probability-sampling mechanism. Instead, we assume for now that the $\mathbf{x}_k \varepsilon_k$ are uncorrelated across PSUs, have bounded variances, and are independent of the first-stage stratification, which is to say the first-stage stratification is ignorable. The sample design is as above except that PSUs could have been selected without replacement. In fact, if there were selected with replacement, no PSU was been selected twice.

Under mild additional asymptotic assumptions, which revolve around the number of sampled PSUs being sufficiently large while the dimension of \mathbf{x}_k is bounded, a design-based variance estimator for \mathbf{b} is

$$\mathbf{var}(\mathbf{b}) = N^{-2} \mathbf{D} \sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{j=1}^{n_h} \left(\sum_{k \in S_{hj}} w_k \mathbf{x}_k e_k - \frac{1}{n_h} \sum_{\varphi=1}^{n_h} \sum_{\kappa \in S_{h\varphi}} w_\kappa \mathbf{x}_\kappa e_\kappa \right) \left(\sum_{k \in S_{hj}} w_k \mathbf{x}_k e_k - \frac{1}{n_h} \sum_{\varphi=1}^{n_h} \sum_{\kappa \in S_{h\varphi}} w_\kappa \mathbf{x}_\kappa e_\kappa \right)^T \mathbf{D}, \quad (6)$$

where H is the number of strata (which need not be bounded), $n_h > 1$ the number of PSUs in stratum h ,

S_{hj} is the set of elements in PSU hj , $\mathbf{D} = \left[N^{-1} \sum_S w_k f'(\mathbf{x}_k^T \mathbf{b}) \mathbf{x}_k \mathbf{x}_k^T \right]^{-1}$, and $e_k = y_k - f(\mathbf{x}_k^T \mathbf{b})$. The ignorability of the first-stage stratification and the independence of the ε_k across PSUs assures the near unbiasedness of the variance estimator in equation (6) under mild conditions as well as the near unbiasedness of

$$\mathbf{var}_A(\mathbf{b}) = \mathbf{D} \sum_{h=1}^H \sum_{j=1}^{n_h} \left(N^{-1} \sum_{k \in S_{hj}} w_k \mathbf{x}_k e_k \right) \left(N^{-1} \sum_{k \in S_{hj}} w_k \mathbf{x}_k e_k \right)^T \mathbf{D}. \quad (7)$$

Under the standard model, the w_k in both equations (5) and (6) can be replaced by $w_k g$.

The key to both variance estimators is that the $\mathbf{E}_{hj} = \sum_{k \in S_{hj}} w_k \mathbf{x}_k \varepsilon_k$ have mean 0 and are uncorrelated across PSUs.

The use of robust sandwich-type variance estimation allows the variance matrices of the \mathbf{E}_{hj} be unspecified. The additional asymptotic assumptions allow $e_k = \varepsilon_k - f'(\theta_k) \mathbf{x}_k^T (\mathbf{b} - \boldsymbol{\beta})$ to be used in place of ε_k within the \mathbf{E}_{hj} , and \mathbf{D} to replace the probability limit of $\mathbf{A}_0 = N^{-1} \sum_S w_k f'(\theta_k) \mathbf{x}_k \mathbf{x}_k^T$.

Additional variations of the variance estimator in equation (5) 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 assumptions that the ε_k are uncorrelated across elements (or another higher-stage of sampling like arousing unit in a household-based sample of individuals) and the more reasonable the assumption that the first-stage stratification is ignorable.

Variance Estimation When First-Stage Stratification is Not Ignorable

When the first-stage stratification is not ignorable, it is tempting to follow design-based theory and argue under probability-sampling theory the \mathbf{E}_{hj} are independent and have a common mean within strata, justifying the use of the variance estimator in equation (6) but not (7). This argument is only valid when the first-stage PSU's are indeed selected with replacement (meaning the same PSU can be selected twice) or their selection can reasonably be approximated by that design. When the first-stage sample is drawn without replacement, strata with few PSUs in both the sample and population can void the near unbiasedness of the variance estimator (invoking large-sample or large-population properties when the corresponding sample or population is not large is dubious). Moreover, as Graubard and Korn (2002) point out, even under with-replacement sampling of PSUs, equation (6) provides a nearly unbiased variance 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 variance of \mathbf{b} that equation (6) fails to capture: the random number of elements within each first-stage stratum when the number of strata is bounded and the mechanism generating the strata when it is not.

Degrees of Freedom

When fitting a regression model with survey, design-based practice often treats the diagonals of the variance estimator in equation (6) 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) 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 multi-dimensional 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 variance of a diagonal of $\mathbf{var}(\mathbf{b}) \approx \mathbf{var}(N^{-1}\sum\sum \mathbf{D}\mathbf{E}_{hj})$, call it r , can be calculated under those assumptions. Using Satterthwaite approximation, the *effective* degrees of freedom for the corresponding component of \mathbf{b} would then be $2/r$ and could vary across coefficients. Although this procedure is itself more than a little dubious, when employed in the generation of t statistics will likely produces better coverages 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.

Tests for Choosing Weights

Suppose an analyst wants to compare 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 modified weights. This would provide an indirect test of the standard model, since using the modified weights produces a nearly unbiased 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 $\mathbf{var}(\cdot)$ is a variance estimator analogous to the one in either equation (4) or (5). A popular design-based test statistic for whether \mathbf{b} and \mathbf{b}' are estimating the same thing is

$$F_{r,d-r+1} = \left(\frac{d-k+1}{d} \right) \frac{(\mathbf{b} - \mathbf{b}')^T [\mathbf{var}(\mathbf{b} - \mathbf{b}')]^{-1} (\mathbf{b} - \mathbf{b}')}{r}, \quad (8)$$

where d is the *nominal* degrees of freedom, that is, $n - H$. The F test in equation (8) is called the *adjusted Wald F*

test in SUDAAN, which also offers a host of variations.

This 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. 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 $(\mathbf{x}_k^T \ \mathbf{0}^T)$ for the first copy and by (\mathbf{x}_k^T) for the second. The regression coefficient is then

$$\mathbf{d} = \begin{pmatrix} \mathbf{b} \\ \mathbf{b}' - \mathbf{b} \end{pmatrix},$$

and testing whether $\mathbf{b}' - \mathbf{b}$ is significantly different from $\mathbf{0}$ becomes a straight-forward regression exercise using design-based software, such as SUDAAN 11 (Research Triangle Institute 2012) and the survey procedures in SAS/STAT 14.1 (SAS Institute Inc. 2015).

A 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).

Another Possible Test for the Standard Model

Here is another test of whether the standard model is consistent with the sampled elements. Estimate \mathbf{b} using inverse-selection-probability weights or modified weights. Compute $f_k = f(\mathbf{x}_k^T \mathbf{b})$. Apply “design-based” 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 because $E(\varepsilon_k | \mathbf{x}_k)$ is clearly not 0 (f_k being a function of \mathbf{x}_k , and the “design-based variance estimator being robust to the heteroscedasticity 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.

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 a design-sensitive model-based frame. Nevertheless, although inserting weights into an estimating equations is often justified, it is not always necessary, depending on what assumptions are made. In addition, although the sandwich-type variance estimator used in design-based practice (equation (6)) may be fairly robust, it does not fully account for first-stage stratification when first-stage stratification is not ignorable. When it is ignorable, a simpler variance estimator (equation (7)) can be used that is likely more stable (i.e., it diagonals have less relative variance). Other, even more stable, variance estimators can be constructed by assuming 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 either package. To do so one treat 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 variance estimators to handle the correlation of the equations. Testing the “parallel lines” assumption of the proportional-odd model that all the $\beta_1 = \beta$ in equation (4) is straightforward.

One interesting repercussion of assuming the standard model is that listwise deletion turns out to be a nearly unbiased 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 variance estimators in equations (6) and (7) were derived using Taylor-series linearization. Replication

techniques such as a jackknife, bootstrap, and often balanced repeated replication will often produce asymptotically equivalent variance estimators. See, for example, Krewski and Rao (1981). Although the probability-sampling proofs in the literature assumes with-replacement sampling within first-stage strata, model-based analogues are straight forward. When comparing weighting methods, the replicates need to be constructed analogously for the two methods.

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 the fit a regression model. Moreover, it can be misleading when survey statisticians graft a finite-sample techniques like degrees of freedom onto what is actually 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 modified weights. Design-based practice has always been to fear that such tests 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 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. A 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.
- 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. & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17, 73-96.
- Kott, P.S. (2007). Clarifying some issues in the regression analysis of survey data. *Survey Research Methods*. 1, 11–18.
- Kott, P.S. (1994). Hypothesis testing of linear regression coefficients with survey data. *Survey Methodology*. 20, 159–164.
- Krewski D. & Rao J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- Korn, E. L. & Graubard, B. I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *American Statistician*, 44, 270-276.
- Lohr, S. (2010). *Sampling: Design and Analysis*, Second Edition, Boston: Brooks/Cole.
- Pfeffermann, D. & Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhya-The Indian Journal of Statistics*, 61(Series B), 166-186.
- Research Triangle Institute (2012). *SUDAAN Language Manual*, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.
- 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.

Wilkinson, L. & 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>).