Survey Research Methods (2007)
http://w4.ub.uni-konstanz.de/srm
Vol. 1, No. 1, pp. 11-18

Clarifying Some Issues in the Regression Analysis of Survey Data

Phillip S. Kott
National Agricultural Statistics Service

The literature offers two distinct reasons for incorporating sample weights into the estimation of linear regression coefficients from a model-based point of view. Either the sample selection is nonignorable or the model is incomplete. The traditional sample-weighted least-squares estimator can be improved upon when the sample selection is nonignorable, but not when the standard linear model fails and needs to be extended.

Conceptually, it can be helpful to view the realized sample as the result of a two-phase process. In the first phase, the finite population is drawn from a hypothetical superpopulation via simple random (cluster) sampling. In the second phase, the actual sample is drawn from the finite population. In the extended model, the parameters of this superpopulation are vague. Mean-squared-error estimation can become problematic when the primary sampling units are drawn within strata using unequal probability sampling without replacement. This remains true even under the standard model when certain aspects of the sample design are nonignorable.

Keywords: Error term; standard linear model; extended linear model; nonignorable; sample design

1 Introduction

How best to estimate the coefficients in a linear model when the observations derive from a sample survey has generated considerable interest in the literature. Kott (1996) provides one model-based argument for incorporating sample weights into the linear regression estimator.

We will restrict our attention here to a *semi-parametric model*, by that we mean a stochastic model where the functional form of the distribution of the error term is not assumed. We will reformulate the argument in Kott within a fully stochastic framework that weakens the *standard linear model* by assuming only that the error term is uncorrelated with the explanatory variables. We will call this the "extended model" (Kott's term, but given a stochastic definition here).

Under the extended linear model, a simple estimating equation leads to an obvious solution: the traditional sample-weighted regression estimator. This is the sample-based analogue to an old result in the econometrics literature (White 1980). Adding the stronger assumption of the standard linear model, $E(\epsilon|\mathbf{x}) = 0$, allows the construction of more efficient estimators.

An alternative rationale for using traditional sample-weighted least squares assumes the standard model holds but allows the error terms to be correlated with the sampling weights. Under this framework, Magee (1998) and Pfeffermann and Sverchkov (1999) show how the sample-weighted least-squares estimator can be improved upon. When measuring the mean squared error of estimated regression coefficients based on a sample from a finite population, it is often helpful to assume the realized sample derives from a two-phase process. In the first phase, the finite population is drawn from a hypothetical superpopulation via simple random (cluster) sampling. In the second phase, the actual sample is drawn from the finite population. Some think that the standard practice of treating the (cluster) sample as if it was drawn with replacement from the finite population is roughly equivalent to the full two-phase process. That is not always the case. Under stratified sampling, the standard practice can miss a component of variance (see Korn and Graubard (1976), although that component will be small when the finite population is large compared to the sample size and can often be "defined away" by conditioning on realized stratum-population fractions. An addition problem is encountered when there are unequal selection probabilities among the primary sampling units within the strata as we will see.

Section 2 lays out the basic framework of the extended and standard linear models. Section 3 provides a simple example of how a zero-meaned error term can be uncorrelated with an explanatory variable but have a mean other than zero when conditioned on it. Section 4 contains some needed asymptotic (both large population and large sample) theory. The notion of a complex random sample is first introduced in Section 5. Section 6 addresses variance estimation, where stratification can have confounding effects. Section 7 discusses how to create a more efficient estimator under the standard model when the data derives from a complex sample. Section 8 extends the previous analysis to a particular class of non-linear models. Section 9 provides some concluding remarks. Our primary goal throughout is conceptual clarity rather than mathematical rigor. Many of the missing proofs can be found by adapting arguments in Binder (1983).

Contact information: National Agricultural Statistics Service, 3251 Old Lee Highway, Fairfax, VA 22030, USA (Phil_Kott@nass.usda.gov).

2 The Framework

Suppose we are interested in estimating the following *extended* linear model describing a relationship among variables in a population:

$$y_i = \mathbf{x}_i\beta + \epsilon_i, \tag{1}$$

where i (= 1, ..., M) denotes an element of the population, $\mathbf{x}_i = (1, \mathbf{z}_i'), \mathbf{z}_i$ is a (p-1)-component vector of variable values associated with element i, β is an unknown p-component column vector, and ϵ_i is a random variable with mean zero.

Our *weak assumption* about the error term ϵ_i in equation (1) is $E(\mathbf{x}_i'\epsilon_i) = \mathbf{0}_p$ for all i. This is much weaker – and thus more general – than the assumption in the *standard* linear model, $E(\epsilon_i|\mathbf{x}_i) = 0$ for every possible $\mathbf{x}_i$. The latter implies that ϵ_i is not only uncorrelated with the components of $\mathbf{x}_i$ but also with any function of the components of $\mathbf{x}_i$.

If every member of the population is an equally likely realization of the model in equation (1), then $E[\sum \mathbf{x}_i'(y_i - \mathbf{x}_i\beta)] = \mathbf{0}_p$. This suggests we estimate β with the vector **b** that satisfies the *estimating equation*, $\sum \mathbf{x}_i'y_i = \sum \mathbf{x}_i'\mathbf{x}_i\mathbf{b}$. A unique solution to this equation exists when $\sum \mathbf{x}_i'\mathbf{x}_i$ is invertible, which we assume to be the case for convenience. That solution is the ordinary least squares (OLS) estimator, $\mathbf{b}_{OLS} = (\sum \mathbf{x}_i'\mathbf{x}_i)^{-1}\sum\mathbf{x}_i'y_i$, which is consistent under the extended model given the asymptotic framework to be described in Section 4. The OLS estimator is not necessarily unbiased if $\mathbf{x}_i$ is a random variable.

The derivation of $\mathbf{b}_{OLS}$ results directly from the weak assumption, $E(\mathbf{x}_i'\epsilon_i) = \mathbf{0}_p$. If, however, we add that $E(\epsilon_i|\mathbf{x}_i) = 0$, then $E[(\mathbf{x}_i'\epsilon_i g(\mathbf{x}_i)] = \mathbf{0}_p$ for any function of $\mathbf{x}_i$. Indeed, suppose $E(\epsilon|\mathbf{X}) = \mathbf{0}_M$, where $\epsilon = (\epsilon_1, ..., \epsilon_M)'$, and **X** is the M x p matrix with $\mathbf{x}_i$ in its i^{th} row. This is a slightly stronger assumption than $E(\epsilon_i|\mathbf{x}_i) = 0$ in principle, but effectively the same in practice. Observe that now $E[\mathbf{X}'\mathbf{G}(\mathbf{X})\,\epsilon] = \mathbf{0}_p$, where **G** in any M x M matrix function of **X**. This last equality suggests the estimating equation:

$$\mathbf{X}'\mathbf{G}(\mathbf{X})(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}_p, \tag{2}$$

where $\mathbf{y} = (y_1, ..., y_M)'$.

It is not hard to see that solving equation (2) for **b** provides a consistent estimator of β; namely, $\mathbf{b}_G = [\mathbf{X}'\mathbf{G}(\mathbf{X})\,\mathbf{X}]^{-1}\mathbf{X}'\mathbf{G}(\mathbf{X})\,\mathbf{y}$. In fact, $\mathbf{b}_G$ is unbiased given **X** under the standard model. Moreover, it is well known that the most efficient solution obtains when **G**(**X**) is related to the variance of ϵ; in particular, when $E(\epsilon\epsilon'|\,\mathbf{X}) = \Omega = [k\mathbf{G}(\mathbf{X})]^{-1}$ for an arbitrary constant k. Consequently, the form given to **G**(**X**) in practice usually reflects some estimate of, or belief about, Ω.

In truth, Ω is rarely known even up to a constant. Throughout the text, we will take the position that one can have some reasonable hypothesis about Ω and incorporate it into the choice of **G**(**X**). Nevertheless, the hypothesis is potentially in error, and variance estimation schemes should protect against that possibility.

In principle, we may want to hypothesize that Ω depends, at least in part, on population variables that are not functions of $\mathbf{x}_i$. The implied extension to the argument of **G**(.) adds nothing substantive to the discussion and will be ignored here.

Notice that when **G**(**X**) in equation (2) is set equal to $\mathbf{I}_M$, the equation becomes the estimating equation for the extended model. Moreover, the equation, $E(\mathbf{X}'\epsilon) = \mathbf{0}_p$ is precisely the weak assumption of the extended model stated in matrix form.

3 An Example

The following example shows why the weaker assumption, $E(\mathbf{x}_i'\epsilon_i) = \mathbf{0}_p$, can be a useful alternative to the standard model assumption, $E(\epsilon_i|\mathbf{x}_i) = 0$. Suppose we have a population in which the relationship $y_i = z_i^{\gamma}$ for $\gamma \neq 1$ is strictly satisfied. We do not know this, however, and try to set $\mathbf{x}_i\beta$ in equation (1) equal to $\beta_1 + \beta_2 z_i$.
The OLS estimates for β_2 and β_1 are

$$b_2 = \frac{\sum z_i^{\gamma+1} - \sum z_i^{\gamma} \sum z_i/M}{\sum z_i^2 - [\sum z_i]^2/M},$$

and

$$b_1 = \sum(z_i^{\gamma} - b_2 z_i)/M,$$

respectively. We make the relatively mild assumption that the series $\sum z_i^{\alpha}/M$ converges to a constant, say $z^{(\alpha)}$, as M grows arbitrarily large, where α can have any of the following values: 1, 2, γ, or γ+1. Under this assumption, b_2 and $b_1$converge to

$$\beta_2 = \frac{z^{(\gamma+1)} - z^{(\gamma)}z^{(1)}}{z^{(2)} - [z^{(1)}]^2},$$

and

$$\beta_1 = z^{(\gamma)} - \beta_2 z^{(1)},$$

respectively. It is now easy to see that

$$\begin{aligned}\epsilon_i &= z_i^{\gamma} - (\beta_1 + \beta_2 z_i)\\ &= (z_i^{\gamma} - z^{(\gamma)}) - \frac{z^{(\gamma+1)} - z^{(\gamma)}z^{(1)}}{z^{(2)} - [z^{(1)}]^2}(z_i - z^{(1)}).\end{aligned}$$

Although ϵ_i can have mean zero and be uncorrelated with z_i, $E(\epsilon_i|z_i)$ can not be equal to zero for all z_i. In fact, ϵ_i is clearly a function of z_i.

The example shows that the weak assumption $E(\mathbf{x}_i'\epsilon_i) = \mathbf{0}_p$ allows a flexibility in model construction that is unavailable with $E(\epsilon_i|\mathbf{x}_i) = 0$. Since reality very seldom fits a postulated model, this flexibility is fortuitous. In our example, when $E(\epsilon_i|\mathbf{x}_i) = 0$ is assumed, the model $y_i = \beta_1 + \beta_2 z_i + \epsilon_i$ is simply wrong, and its parameters cannot be estimated. When $E(\mathbf{x}_i'\epsilon_i) = \mathbf{0}_2$ is assumed, however, the parameters of the model *can* be estimated. Many will argue that we should not estimate parameters for "'wrong"' models, but aren't all models wrong?

4 Some Asymptotic Theory

In this section, we develop some theory for $\mathbf{b}_{\mathrm{G}}$ under the standard linear model. It is straightforward to do the same for $\mathbf{b}_{\mathrm{OLS}}$ under the extended model by setting $\mathbf{G}(\mathbf{X})$ equal to $\mathbf{I}_{\mathrm{M}}$.

For $\mathbf{b}_{\mathrm{G}}$ to be a consistent estimator for β, a number of asymptotic conditions must be satisfied. It is sufficient that as M grows arbitrarily large

$$\lim_{M\to\infty} (\mathbf{X}'\mathbf{G}(\mathbf{X})\mathbf{X}/M) = \mathbf{F}, \tag{3.1}$$

and

$$\operatorname*{plim}_{M\to\infty} (\mathbf{X}'\mathbf{G}(\mathbf{X})\epsilon/ \rfloor\, M) = \mathbf{d}, \tag{3.2}$$

for some positive definite matrix $\mathbf{F}$ and bounded vector $\mathbf{d}$. Under these conditions – referred to collectively as equation (3), one can easily show that $\mathbf{b}_{\mathrm{G}} - \beta = \mathbf{O}_{\mathrm{p}}(\rfloor[1/\mathrm{M}])$. These assumption do no require $\mathbf{G}(\mathbf{X})$ to have a particular form. It should prove helpful, however, for only $\mathrm{O}_{\mathrm{P}}(\mathrm{M})$ of the terms in the M x M matrix to be non-zero, so that each entry of the matrix $\mathbf{X}'\mathbf{G}(\mathbf{X})\ \mathbf{X}$ would be the sum of no more than $\mathrm{O}_{\mathrm{P}}(\mathrm{M})$ terms.

Let $\mathbf{D}_{\mathrm{i}}$ be an M x M with zeroes everywhere but the i^{th} diagonal, which is 1. Let $\mathbf{u}_{\mathrm{i}} = \mathbf{X}'\mathbf{G}(\mathbf{X})\ \mathbf{D}'_{\mathrm{i}}\epsilon$ (this vector depends on the choice for $\mathbf{G}(.)$, but we suppress that in the notation). Suppose the population can be grouped into J mutually exclusive *clusters*, denoted C(1), ..., C(J), such that $\mathrm{E}(\mathbf{u}_{\mathrm{i}}\mathbf{u}'_{\mathrm{k}}|\mathbf{X})$ is non-negative definite when i and k are in the same cluster and equal to $\mathbf{0}_{\mathrm{pxp}}$ otherwise. An analogous assumption about $\mathrm{E}(\mathbf{u}_{\mathrm{i}}\mathbf{u}'_{\mathrm{k}})$ is needed under the extended model. Although we will relax many assumptions in this discussion, we will not allow the ones above to fail, at least not when variances need to be estimated.

In many practical situations, the M elements in the population will serve as the J clusters. In others, there will be a clear need to collect elements whose error terms can not be assumed uncorrelated into clusters, as we shall see in the following section.

In practice, a good choice for $\mathbf{G}(\mathbf{X})$ will mimic the cluster structure. That is to say $\mathrm{G}_{\mathrm{ik}}(\mathbf{X})$ will be zero when i and k are in different clusters. We will assume that to be the case for the remainder of the text. Moreover, we will assume $\mathbf{G}$ to be symmetric.

The "sandwich estimator" for the variance of $\mathbf{b}_{\mathrm{G}}$ is

$$\mathbf{V} = [\mathbf{X}'\mathbf{G}(\mathbf{X})\mathbf{X}]^{-1} \sum_{j=1}^{J} \mathbf{R}_{j+}\mathbf{R}'_{j+}[\mathbf{X}'\mathbf{G}(\mathbf{X})\mathbf{X}]^{-1} \tag{4}$$

where $\mathbf{R}_{\mathrm{j+}} = \mathbf{X}'\mathbf{G}(\mathbf{X})\ \mathbf{D}_{\mathrm{j+}}(\mathbf{y} - \mathbf{X}\mathbf{b}_{\mathrm{G}})$, and $\mathbf{D}_{\mathrm{j+}}$ is a diagonal matrix with 1's in the rows corresponding to elements of cluster C(j) and 0's everywhere else. Note that $\mathbf{R}_{\mathrm{j+}}$ approximately equals the sum of the $\mathbf{u}_{\mathrm{i}}$ across the elements in C(j). If J/M converges to a positive constant as M grows arbitrarily large, then it is not difficult to show under the assumptions we have made that $\mathbf{V}$ is an asymptotically unbiased estimator for the variance of $\mathbf{b}_{\mathrm{G}}$.

5 Random Sampling

Solving equation (2) to derive an estimator for β assumes that the M elements in the population are generated by a process that produces elements satisfying equation (1). Moreover, were the process allowed to continue, the two parts of equation (3) would likewise be satisfied.

Following Fuller (1975), we will treat the J clusters in the population as if they were a simple random sample from a putative infinite population, each of whose elements satisfy equation (1). Moreover, as the number of these clusters (and therefore M) grows arbitrarily large, equation (3) continues to hold.

We are now ready to address the main concern of this paper. Sometimes analysts do not have access to information on all the variables in equation (1) for the entire population. Instead, a probability sample is drawn, and a complete set of variable values is collected only for the sample. We will concentrate here on a stratified, multi-stage sample and ignore the possibility of element or item (i.e., variable) nonresponse.

Suppose that, before sampling, the J clusters in the population are separated into H mutually exclusive strata (H may be 1). A probability sample of n_{h} clusters is selected within each stratum h *without* replacement (from now on, all samples are assumed to be drawn without replacement). The n $= \sum \mathrm{n}_{\mathrm{h}}$ sampled clusters are called primary sampling units (PSU's). Probability samples of elements are drawn independently within each PSU. We allow the possibility that all the elements in a PSU are drawn into the sample or that the PSU's are themselves elements. Let S denote the element sample and m be the size of S.

If $\mathrm{E}(\mathbf{u}_{\mathrm{i}}|\mathbf{X};\ \mathrm{i} \in \mathrm{S}) = \mathbf{0}_{\mathrm{p}}$, then solving $\mathbf{X}'\mathbf{D}_{\mathrm{S}}\mathbf{G}(\mathbf{X})\mathbf{D}_{\mathrm{S}}(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}_{\mathrm{p}}$ for $\mathbf{b}$, where $\mathbf{D}_{\mathrm{S}} = \sum_{\mathrm{S}}\mathbf{D}_{\mathrm{i}}$, provides an unbiased and consistent estimator for β under mild conditions. The assumption that $\mathrm{E}(\mathbf{u}_{\mathrm{i}}|\mathbf{X};\ \mathrm{i} \in \mathrm{S}) = \mathbf{0}_{\mathrm{p}}$, that sample selection is *ignorable*, effectively means that there is no information about y_{i} in the element selection probabilities not captured by $\mathbf{x}_{\mathrm{i}}\beta$.

What if that is not the case?

The solution is well-known in the literature on randomization-based inference (see, for example, Binder 1983). Let t_{i} be the *sample-inclusion indicator for* element i,a random variable equal to 1 when i ∈ S and 0 otherwise. Furthermore, let π_{i} be the selection probability of i (i.e., $\mathrm{E}(\mathrm{t}_{\mathrm{i}}) = \pi_{\mathrm{i}}$), and w_{i} be its *sample weight*, $1/\pi_{\mathrm{i}}$. Call the M x M diagonal matrices with t_{i} and w_{i} in their i^{th} position, $\mathbf{T}$ and $\mathbf{W}$, respectively. (Note that after the sample is drawn, $\mathbf{T} = \mathbf{D}_{\mathrm{S}}$.)

Suppose the vector $\mathbf{b}_{\mathrm{wG}}$ solves the equation:

$$\mathbf{X}'\mathbf{G}(\mathbf{X})\mathbf{T}\mathbf{W}(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}_{\mathrm{p}}. \tag{2'}$$

When $\mathbf{X}'\mathbf{G}(\mathbf{X})\mathbf{T}\mathbf{W}\mathbf{X}$ is invertible, $\mathbf{b}_{\mathrm{wG}} = [\mathbf{X}'\mathbf{G}(\mathbf{X})\ \mathbf{T}\mathbf{W}\mathbf{X}\ ]^{-1}\mathbf{X}'\mathbf{G}(\mathbf{X})\ \mathbf{T}\mathbf{W}\mathbf{y}$. Applying equation (2') in its most general form requires knowledge of the complete $\mathbf{X}$ matrix. That will often not be the case in practice, but there remains a host of viable choices for $\mathbf{G}(\mathbf{X})$. We return to the issue of choosing $\mathbf{G}(\mathbf{X})$ in Section 7.

If we assume that the variables and sample design are such that

$$\lim_{n\to\infty}\left(\sum \mathbf{X'G(X)TWX}/M\right) = \mathbf{F}_{\mathrm{w}}, \tag{3.1'}$$

and

$$\operatorname*{plim}_{n\to\infty}\left(\sum \mathbf{X'G(X)TW}\epsilon/\rfloor\, M\right) = \mathbf{d}_{\mathrm{w}}, \tag{3.2'}$$

for some invertible matrix $\mathbf{F}_{\mathrm{w}}$ and bounded vector $\mathbf{d}_{\mathrm{w}}$, then under mild conditions, $\mathbf{b}_{\mathrm{wG}} - \beta = \mathbf{O}_{\mathrm{p}}(\rfloor[1/n])$. Note, however, that it is possible for $\mathbf{b}_{\mathrm{wG}} - \beta$ conditioned on a particular sample to *not* approximately equal zero. The near asymptotic unbiasedness of $\mathbf{b}_{\mathrm{wG}}$ occurs when we average over all possible samples.

In order to apply equation (3'), we need to impose an asymptotic framework on the sample design. We do this by assuming an infinite sequence of samples and populations, $\{S_\nu\}$ and $\{P_\nu\}$. Let m_ν denote the number of elements in S_ν, M_ν the analogous size of P_ν, n_ν the number of PSU's in S_ν, J_ν the number of clusters in P_ν, H_ν the number of strata in both S_ν and P_ν, and $n_{h\nu}$ the number of PSU's from stratum h in S_ν.

As ν grows arbitrarily large, so does n_ν. The ratios, m_ν/n_ν, J_ν/n_ν, and M_ν/m_ν all converge to positive constants. When H is small, it makes sense to assume an asymptotic framework in which H_ν stays the same as ν grows, and the $n_{h\nu}/n_\nu$ converge to positive constants. Otherwise, the $n_{h\nu}$ can be assumed stay the same while H_ν/n_ν converges to a positive constant *It is important to realize that full-population equation (3) with M_ν replacing M is assumed to hold for each value of ν.*

6 Variance Estimation for $\mathbf{b}_{\mathbf{wG}}$

6.1 Ideal circumstance

Suppose $E(\mathbf{u}_i\mathbf{u}_k'|\mathbf{X}) = E(\mathbf{u}_i\mathbf{u}_k'|\mathbf{X};\, i, k \in S) = \mathbf{0}_{\mathrm{pxp}}$ when i and k are from different clusters and non-negative definite otherwise (for the extended model, replace $E(\mathbf{u}_i\mathbf{u}_k'|\mathbf{X};\ .\ )$ with $E(\mathbf{u}_i\mathbf{u}_k';\ .\ )$). Under mild conditions, we can estimate the variance/mean squared error of $\mathbf{b}_{\mathrm{wG}}$ with an analogue of $\mathbf{V}$ in equation (4), namely,

$$\mathbf{V}^* = [\mathbf{X'G(X)TWX}]^{-1}\sum_{j=1}^{J}\mathbf{R}_{wj+}\mathbf{R}_{wj+}'[\mathbf{X'TWG(X)X}]^{-1} \tag{5}$$

where $\mathbf{R}_{\mathrm{wj+}} = \mathbf{X'G(X)\ TWD}_{\mathrm{j+}}(\mathbf{y}_{\mathrm{j+}} - \mathbf{X}_{\mathrm{j+}}\mathbf{b}_{\mathrm{wG}})$. In many applications, however, $E(\mathbf{u}_i\mathbf{u}_k'|\mathbf{X};\, i, k \in S)$ may not equal $\mathbf{0}_{\mathrm{pxp}}$ when i and k are from different clusters.

6.2 Stratification

Before discussing a more general variance or mean-squared-error estimator (the former expression is used exclusively from here on), we first investigate the potential effect of stratification on estimation under the extended model. Analogous arguments can be made for the standard model.

Let Γ_{i} be an integer from 1 to H indicating which stratum contains i. It is tempting to assume that stratification is ignorable in the sense that $E(\mathbf{u}_i|\,\Gamma_i = h) = \mathbf{0}_{\mathrm{p}}$ (heuristically, there is no additional information in the stratification not already captured by the model). This assumption is problematic in some applications, especially those with a large number of strata. The more general assumption that $E(\mathbf{u}_i|\,\Gamma_i = h) = \mathbf{q}_{\mathrm{h}}$ appears to be an attractive alternative, but it is not very helpful under multi-stage sampling when the implicit estimate of the population size for a stratum is random.

Instead, we will again adapt the pragmatic approach of invoking the randomization-based properties of the estimator. That is to say, we will treat the $\mathrm{t_i}$ as random variables rather than conditioning on the realized sample − the usual practice in most of statistics, but not survey sampling. Mathematically, this changes the goal of variance estimation from $E[(\mathbf{b}_{\mathrm{G}} - \beta)(\mathbf{b}_{\mathrm{G}} - \beta)'|S]$ for every S to $E[(\mathbf{b}_{\mathrm{G}} - \beta)(\mathbf{b}_{\mathrm{G}} - \beta)']$.

Let $\mathbf{u}_{\mathrm{hj+}}$ be the sum of the $\mathbf{u}_{\mathrm{i}}$ ($\approx \mathbf{X'G(X)D}_{\mathrm{i}}\epsilon$) across all the subsampled elements in PSU j of stratum h. The expectation of $\mathbf{u}_{\mathrm{hj+}}$ is constant across the sampled PSU's within a stratum, but can vary across strata. Consequently, we need to assume that N_h/N, where N_h is the number of clusters in stratum h, stays constant as $N = \sum^H N_h$, and M grow arbitrarily large in equations (3) and (3'). This means that the fraction of the population clusters in each stratum does not change as the population grows arbitrarily large. If it did, there could be another component of variance not captured by the variance estimator to be discussed below. For more on the missing variance component, see Korn and Graubard (1976).

The randomization-based variance estimator for $\mathbf{b}_{\mathrm{wG}}$ (see, for example, Binder (1983)) is

$$\mathbf{V}_{RB} = [\mathbf{X'G(X)TWX}]^{-1}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{j\in S_h^*}\mathbf{R}_{wj+}\mathbf{R}_{wj+}' - n_h^{-1}$$
$$\left(\sum_{j\in S_h^*}\mathbf{R}_{wj+}\right)\left(\sum_{j\in S_h^*}\mathbf{R}_{wj+}\right)'[\mathbf{X'TWG(X)X}]^{-1} \tag{6}$$

where S_h^* is the set of sampled PSU's in stratum h. It is asymptotically unbiased under mild conditions when PSU's are selected using stratified, simple random sampling conditioned on the realized N_h/N values. This is because within each stratum h, the $\mathbf{R}_{\mathrm{wj+}}'$s selected for the sample are estimates of the same value and asymptotically uncorrelated.

Observe that when H =1 in equation (6), $\mathbf{V}_{\mathrm{RB}}$ will become asymptotically indistinguishable from $\mathbf{V}^*$ in equation (5) when $\sum_{S^*}\mathbf{R}_{\mathrm{wj+}} = \mathbf{0}_{\mathrm{p}}$, as happens when $\mathbf{G}$ is the identity matrix.

If $E(\mathbf{u}_i\mathbf{u}_k'|\mathbf{X};\, i, k \in S) = \mathbf{0}_{\mathrm{pxp}}$ when i and k are from different clusters, then $\mathbf{V}_{\mathrm{RB}}$ can easily be shown to be, like $\mathbf{V}^*$, asymptotically unbiased given any sample. It is not as efficient, however.

6.3 *Unequal probability sampling*

In most practical situations, $\mathbf{V}_{\mathrm{RB}}$ will be reasonable – although not necessarily asymptotically unbiased – when PSU's are selected with unequal probabilities within strata. For simplicity, we restrict our attention to a single-stage sample in this section.

Denoting $\mathrm{E(t_i t_k)}$ as π_{ik}, observe that $\mathrm{E}(\epsilon_{\mathrm{i}}\epsilon_{\mathrm{k}}|\mathbf{X}) = 0$ for i ≠ k does not imply $\mathrm{E}(\mathrm{w_i t_i}\epsilon_{\mathrm{i}}\mathrm{w_k t_k}\epsilon_{\mathrm{k}}|\mathbf{X}) = \mathrm{E}(\epsilon_{\mathrm{i}}\epsilon_{\mathrm{k}}\pi_{\mathrm{ik}}/[\pi_{\mathrm{i}}\pi_{\mathrm{k}}]|\mathbf{X}) = 0$ when $\pi_{\mathrm{ik}}/[\pi_{\mathrm{i}}\pi_{\mathrm{k}}]$ varies across sampled elements. That is to say, the joint inclusion indicators may be nonignorable for some designs. This can cause a problem in variance estimation when conditioned on a realized sample, but that problem vanishes when averaging across all possible samples as we did in the previous subsection.

Again for simplicity, consider the simplest version of the model in equation (1), where $\mathbf{x}_{\mathrm{i}} = \mathrm{x_i} = 1$. The model parameter of interest is the scalar β. Given an unclustered population of size M, the full-population estimator for β is $\mathrm{b} = \sum_{\mathrm{U}} \mathrm{y_j/M}$. Now given a probability element sample S, the conventional sampled-weighted estimator for β is

$$b_{\mathrm{w}} = \frac{\sum_S \frac{y_j}{\pi_j}}{\sum_S \frac{1}{\pi_j}}$$

Suppose the M elements in the finite population were chosen using without-replacement simple random sampling from a hypothetical superpopulation of size M^*, which we allow to grow arbitrarily large. The sample itself becomes the second of a two-phase sampling process.

Let each unit j have (second-phase) selection probability π_{j}, and assume that no joint selection probability, $\pi_{\mathrm{ij},}$is zero (which rules out systematic sampling). We are interested here in the (randomization) mean squared error of

$$b^*_w = \frac{\sum \frac{y_j}{\pi(M/M^*)}}{\sum \frac{1}{\pi_j(M/M^*)}}$$

as M^* grows arbitrarily large.

We can rewrite

$$b^*_w \text{ as } \beta + \frac{\sum \frac{\epsilon_{\mathrm{j}}}{\pi_{\mathrm{j}}(M/M^*)}}{\sum \frac{1}{\pi_{\mathrm{j}}(M/M^*)}} \approx \beta + (M^*)^{-1} \sum \frac{\epsilon_{\mathrm{j}}}{\pi_{\mathrm{j}}(M/M^*)}$$

under the conditions in equation (3'), which we assume to hold. Särndal et al. (Särndal 1992:348, equation 9.3.7) provide an unbiased mean-squared-error estimator for the two-phase estimator $\mathrm{b}^*_{\mathrm{w}}$ given a population of size M^*:

$$var(b^*_w; M^*) = M^{-2}\left(\frac{1-M}{M^*}\right)\sum_{j\in S}\frac{\epsilon_j^2}{\pi_j} - (M-1)^{-1}M^{-2}$$

$$\left(\frac{1-M}{M^*}\right)\sum_{\substack{j\neq k\\ j,k\in S}}\frac{\epsilon_j\epsilon_k}{\pi_{jk}}\sum_{j\in S}\frac{(1-\pi_j)\epsilon_j^2}{\pi_j^2} + M^{-2}\sum_{\substack{j\neq k\\ j,k\in S}}\frac{(\frac{1-\pi_j\pi_k}{\pi_{jk}})\epsilon_j\epsilon_k}{\pi_j\pi_k}$$

where S denotes the sample which has size n. Taking the limit as M^* grows arbitrarily large, and rearranging terms yields the estimator:

$$var(b^*_w; \infty) = v_0 + A \qquad (7)$$

where

$$v_0 = M^{-2}\sum_{j\in S}\frac{\epsilon_j^2}{\pi_j^2} - M^{-2}(n-1)^{-1}\sum_{\substack{j\neq k\\ j,k\in S}}\left(\frac{\epsilon_j}{\pi_j}\right)\left(\frac{\epsilon_k}{\pi_k}\right)$$

$$= M^{-2}\frac{n}{n-1}\left[\sum\left(\frac{\epsilon_j}{\pi_j}\right)^2 - n^{-1}\left(\sum\frac{\epsilon_j}{\pi_j}\right)^2\right]$$

and

$$A = M^{-2}\frac{n}{n-1}\sum_{\substack{j\neq k\\ j,k\in S}}\left(\frac{\epsilon_j\epsilon_k}{\pi_{jk}}\right)\left\{\frac{\pi_{jk}}{\pi_j\pi_k} - (n-1)\frac{M}{n(M-1)}\right\}$$

When the expectation (under the model and across all possible samples) of A in equation (7) is zero, then mean squared error of $\mathrm{b}^*_{\mathrm{w}}$ can be estimated with $\mathrm{v_0}$; that is, as if the sample had been drawn with-replacement from an unstratified finite population. It is not hard to see that if all the unknown ϵ_{i} in $\mathrm{v_0}$ were replaced with their sample analogues ($\mathrm{e_i = y_i - b_W}$), then the resulting variance estimator would be a special case of $\mathbf{V}^*$ in equation (5).

Now

$$E(A) = M^{-2}\frac{n}{n-1}\sum_{\substack{j\neq k\\ j,k\in U}}E(\epsilon_j\epsilon_k)\left\{\frac{\pi_{jk}}{\pi_j\pi_k} - (n-1)\frac{M}{n(M-1)}\right\}$$

is zero under simple random sampling without replacement. It is also zero when $\mathrm{E}(\epsilon_{\mathrm{j}}\epsilon_{\mathrm{k}}) = 0$.

When is $\mathrm{E}(\epsilon_{\mathrm{j}}\epsilon_{\mathrm{k}})$, j≠k, not zero? When the stratification is nonignorable so that $\mathrm{E}(\epsilon_{\mathrm{j}} \mid \Gamma_{\mathrm{j}} = \mathrm{h}) \neq 0$, it is unlikely $\mathrm{E}(\epsilon_{\mathrm{j}}\epsilon_{\mathrm{k}} | \Gamma_{\mathrm{j}} = \Gamma_{\mathrm{k}} = \mathrm{h})$ will be zero. Consider a stratified sampling scheme, where $\mathrm{S_h}$ denotes stratum h (h = 1, ..., H) containing $\mathrm{M_h}$ population and $\mathrm{n_h}$ sample units. If as M^* grows arbitrarily large the relative stratum sizes of the hypothetical superpopulation, the $\mathrm{M}^*_{\mathrm{h}}$, grow in proportion, then it is not hard to see components of equation (7) can be reformulated as

$$var(b^*_w; \infty) = v_{ST} + A_{ST},$$

where

$$v_{ST} = M^{-2}\sum_{h=1}^{H}\left\{\sum_{j\in S_h}\frac{\epsilon_j^2}{\pi_j^2} - (n_h-1)^{-1}\sum_{\substack{j\neq k\\ j,k\in S_h}}\frac{\epsilon_j}{\pi_j}\frac{\epsilon_k}{\pi_k}\right\}$$

$$= M^{-2}\sum\left\{\frac{n_h}{n_h-1}\left[\sum\left(\frac{\epsilon_j}{\pi_j}\right)^2 - n_h^{-1}\left(\sum\frac{\epsilon_j}{\pi_j}\right)^2\right]\right\}$$

and

$$A_{ST} = M^{-2}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{\substack{j\neq k\\ j,k\in S_h}}\frac{\epsilon_j\epsilon_k}{\pi_{jk}}\left\{\frac{\pi_{jk}}{\pi_j\pi_k}-(n_h-1)\frac{M_h}{n_h(M_h-1)}\right\}$$

If all the unknown ϵ_i in v_{ST} were replaced with their sample analogues, then the resulting variance estimator would be a special case of $\mathbf{V}_{WR}$ in equation (6). Observe that $E(A_{ST}) = 0$ when there is simple random sampling within strata, but the equality does not necessarily hold otherwise unless $E(\epsilon_j\epsilon_k|\ \Gamma_j = \Gamma_k = h) = 0$ is as well. Nevertheless, the expectation of the strictly nonnegative v_{ST} is likely to dominate the ambiguously-signed $E(A_{ST})$ in practice. Moreover, under several popular unequal-probability-of-selection designs, $\pi_{jk}/(\pi_j\pi_k) - (n_h - 1)M_h/[n_h(M_h - 1)$ in $E(A_{ST})$ will be approximately zero within strata where $M_h >> n_h$ and no π_i is too large. See Asok and Sukhatme (1976) for a discussion of Sampford sampling and systematic unequal probability sampling from a randomly-ordered list.

7 Choosing G(X)

Under the extended model, $\mathbf{G}(\mathbf{X})$ is set equal to $\mathbf{I}_M$. The resulting estimator, $\mathbf{b}_{wLS}$, is called the "sample-weighted least squares" solution. Under the standard model, however, we are free to choose $\mathbf{G}(\mathbf{X})$ to minimize the mean squared error of $\mathbf{b}_{wG}$. When the w_i are not all equal for i ϵ S, the choice is not obvious even when $E(\epsilon\epsilon')$ is known up to a constant. Moreover, we are usually constrained in practice to $\mathbf{G}(.)$ that are functions only of the $\mathbf{x}_i$ in the sample.

One obvious viable choice for $\mathbf{G}(.)$ is a diagonal matrix with $g(\mathbf{x}_i)$ in the i^{th} diagonal. Magee (1998) considers the case where $E(\epsilon\epsilon') = \Omega$ has an unknown diagonal matrix. He proposes using a quasi-Aitken procedure to choose g(.) from among a family of functions of the form $g(\mathbf{x}_i;\ \alpha)$ (note: in a quasi-Aitken procedure, one chooses α seeking to minimize the estimated variances of the components of $\mathbf{b}_{wG}$ directly rather than through an estimate of Ω). It is unclear how to generalize this procedure when $E(\epsilon\epsilon')$ is not diagonal, however.

On the surface, Pfeffermann and Sverchkov (1999) address an even simpler situation: the case where $\Omega = \sigma^2\mathbf{I}_M$. They suggest setting each $g(\mathbf{x}_i)$ equal to the inverse of an estimate for $E(w_i|\mathbf{x}_i;\ i \in S)$. Effectively, they propose "filtering out" from the sampling weight w_i that part explainable by $\mathbf{x}_i$ (their method of arriving at this proposal is much different from ours, but that need not concern us here). With their approach, generalization to more complex Ω appears straightforward. Let $\mathbf{E}(\mathbf{X})$ be an estimate of, or belief about, Ω up to a constant. Furthermore, let $\mathbf{H}(\mathbf{X})$ be a diagonal matrix with an estimate for $E(w_i|\mathbf{x}_i\ ;\ i \in S)$ in the i^{th} diagonal. Then $\mathbf{G}(\mathbf{X})$ can be set equal to $[\mathbf{E}(\mathbf{X})\mathbf{H}(\mathbf{X})]^{-1}$.

Problems remain, however. If $\mathbf{x}_f$ is unknown when f is not in the sample, we need to replace

$$\mathbf{b}_{wG} = [\mathbf{X}'\mathbf{G}(\mathbf{X})\mathbf{TWX}]^{-1}\mathbf{X}'\mathbf{G}(\mathbf{X})\mathbf{TWy}$$

with something like

$$\mathbf{b}_{wG^*} = [\mathbf{X}'\mathbf{TG}(\mathbf{X})\mathbf{TWX}]^{-1}\mathbf{X}'\mathbf{TG}(\mathbf{X})\mathbf{TWy}.$$

At the heart of this is replacing $\mathbf{G}(\mathbf{X})\mathbf{TW}$ with $\mathbf{T}\ \mathbf{G}(\mathbf{X})\mathbf{TW}$. The former has an expectation equal to $\mathbf{G}(\mathbf{X})$, a function of $\mathbf{X}$. The expectation of the row-i-column-k component of the latter is $G_{ik}(\mathbf{X})E(t_k|\ i \in S)$. This is a function of $\mathbf{X}$ only when $E(t_k|\ i \in S)$ is. Moreover, for $\mathbf{b}_{wG*}$ to be viable, $G_{ik}(\mathbf{X})$ with i and k in S must be a function of only those $\mathbf{x}_f$ with f ∈ S. Finally, $\mathbf{b}_{wG*}$ need not be optimal in any sense even when $\mathbf{b}_{wG}$ would be unless the $E(t_k|\ i \in S)$ are all constant.

A tempting alternative to computing a general $\mathbf{b}_{wG*}$ is to choose a diagonal $\mathbf{G}(.)$, based on the (assumed) diagonals of $\mathbf{E}(.)$ and suffer the loss of efficiency that may imply. No matter how $\mathbf{b}_{wG*}$ is computed, if it is a consistent estimator for β, its variance can be estimated using equation (6) with $\mathbf{TG}(.)$ replacing $\mathbf{G}(.)$ everywhere including the definition of the $\mathbf{R}_{wj+}$:

$$\mathbf{V}_{RB} = [\mathbf{X}'\mathbf{TG}(\mathbf{X})\mathbf{TWX}]^{-1}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{j\in S_h^*}\mathbf{R}_{wj+}\mathbf{R}'_{wj+} - n_h^{-1}$$
$$\left(\sum_{j\in S_h^*}\mathbf{R}_{wj+}\right)\left(\sum_{j\in S_h^*}\mathbf{R}_{wj+}\right)'[\mathbf{X}'\mathbf{TWG}(\mathbf{X})\mathbf{TX}]^{-1} \quad (6')$$

where $\mathbf{R}_{wj+} = \mathbf{X}'\mathbf{T}\ \mathbf{G}(\mathbf{X})\ \mathbf{TWD}_{j+}(\mathbf{y}_{j+} - \mathbf{X}_{j+}\mathbf{b}_{wG})$.

Returning to a diagonal $\mathbf{G}(.)$, either because the sample is single-stage or because the hypothesis of a diagonal Ω seems plausible with the data at hand, Magee's approach is the more straightforward, but it is not clear where the functional form, $g(\mathbf{x}_i;\ \alpha)$, is supposed to come from. Magee provides an empirical example where the "wrong" choice does not hurt very much; that is to say, his method is nearly unbiased and much more efficient than sample-weighted least squares.

One appealing attribute of the Pfeffermann-Sverchkov approach is that the sampling weights are shown to be ignorable when each w_i can be fully expressed as a function of $\mathbf{x}_i$. A practical example of this is when selection probabilities are proportional to z_{i1}, the first component of $\mathbf{z}_i$.

When selection probabilities are proportional to the y_i, it makes sense to estimate $E(w_i|\mathbf{x}_i;\ i \in S)$ up to constant by $\mathbf{x}_i\mathbf{b}_{wLS.}$ Similarly, when the ϵ_i are uncorrelated, and $E(\epsilon_i\ ^2)$ is thought to be proportional to $(\mathbf{x}_i\beta)^{\alpha}$, one can replace the unknown β by $\mathbf{b}_{wLS}$, and use an Aitken or quasi-Aitken technique to choose α. An iterative scheme may return an even more efficient estimator.

8 Nonlinear Models

In this section, we consider the following mild generalization of the model in equation (1):

$$y_1 = f(\mathbf{x}_i\beta) + \epsilon_{\mathbf{i}}, \quad (8)$$

where f is a monotonic, twice-differentiable function. An estimating equation for β under the standard model assumption, $E(\epsilon| \mathbf{X}) = \mathbf{0}_M$, is

$$\mathbf{X}'\mathbf{G}(\mathbf{X})[\mathbf{y} - \mathbf{f}(\mathbf{X}\beta)] = \mathbf{0_p}, \tag{9}$$

where $\mathbf{f}(\mathbf{X}\beta) = (f(\mathbf{x}_1\beta), ..., f(\mathbf{x}_M\beta))'$. We again call the solution to equation (9) $\mathbf{b}_G$.

Under the extended model, where only $E(\mathbf{X}'\epsilon) = \mathbf{0}_p$ is assumed, the arbitrary $\mathbf{G}(\mathbf{X})$ is replaced by $\mathbf{I}_M$. Alternatively, choosing $\mathbf{G}(\mathbf{X})$ proportional to $\mathbf{F}(\mathbf{X})\Omega^{-1}$, where $\mathbf{F}(\mathbf{X})$ is the M x M diagonal matrix with $\partial f(\mathbf{x}_i\beta)/\partial(\mathbf{x}_i\beta)$ in the i^{th} diagonal, minimizes the objective function: $[\mathbf{y} - \mathbf{f}(\mathbf{X}\beta)]'\Omega^{-1}[\mathbf{y} - \mathbf{f}(\mathbf{X}\beta)]$. For the special case of logistic regression with an unclustered population, the best choice for $\mathbf{G}(\mathbf{X})$ under this criterion is simply $\mathbf{I}_M$. The choice for $\mathbf{b}$ under the standard and extended models coincide.

The rest of the analysis of estimating the model in equation (8) closely follows that of the linear models in the previous sections. The big difference is that the $[\mathbf{X}'\mathbf{TG}(\mathbf{X})\mathbf{TWX}]^{-1}$ in equation (6') gets replaced by $[\mathbf{X}'\mathbf{TG}(\mathbf{X})\mathbf{TWF}(\mathbf{X})\mathbf{X}\ ]^{-1}$.

9 Some Concluding Remarks

In the conventional study of linear models, one usually suppresses concern with the sampling mechanism and concentrates entirely on the stochastic nature of the model. In survey sampling, the reverse is often true: the model is suppressed and attention is directed exclusively at the sampling mechanism. The question then is what is being estimated?

Fuller (1975) may have been the first to describe how one can estimate the parameters of a linear model without actually assuming the model. He concedes that an unknown model may have generated the data in the finite population, but he is loath to say much about that model. Kott (1991) attempts to flesh out that model. Borrowing from White (1980), we have put that attempt into a fully stochastic framework here, calling it "the extended model." The key is that if one starts with the linear model in equation (1) and assumes only that ϵ_i and $\mathbf{x}_i$ are uncorrelated, then the parameter β is estimable in most situation because we have an estimating equation with p equations and p unknowns.

The stronger assumption that $E(\epsilon_i| \mathbf{x}_i) = 0$ is what can easily fail in the standard linear model. Adding it allows one to construct a more efficient estimator for β than results from ordinary least squares. Without it, many argue we have no model to estimate at all.

Magee (1998) and Pfeffermann and Sverchkov (1999) address how to efficiently estimate the parameters of a standard linear model when the sampling mechanism can *not* be ignored; in particular, when t_i is correlated with ϵ_i. It is difficult to extend their results to situations where the element errors are clustered, as we have seen.

Both Magee and Pfeffermann and Sverchkov discuss variance estimation, but their results most easily apply under Poisson sampling. Stratification and unequal probability sampling complicate matters, as we have seen.

Several additional points about variance estimation need to be emphasized. When the sampling mechanism is ignorable (effectively that t_i and ϵ_i are uncorrelated and $E(\epsilon_i| \Gamma_i = h) = 0$), $\mathbf{V}_{RB}$ in equation (6) is an unbiased estimator for the variance of $\mathbf{b}_{wG}$, which is also unbiased, conditioned on the realized sample. When $E(\epsilon_i| \Gamma_i = h) \neq 0$ is a possibility, one can abandon conditioning on the realized sample and draw inferences about the hypothetical superpopulation using the randomization-based methods described in the text (equations (6) and (6')). These methods average over all possible samples. Their use is appropriate when the finite- population size is very large compared to the sample size or when one condition on the relative stratum sizes as the superpopulation grows arbitrarily large.

There are situations where such conditioning makes sense, for example, when the stratum divisions are politically-determined regions. The other extreme is when strata are determined using the y_i values. This can happen when the y_i are known for the entire population, and sampling is only needed for the collection of corresponding $\mathbf{x}_i$ values. In that situation, one will often need to measure the additional variance component discussed in Korn and Graubard (1976). Graubard and Korn (2002) explores this matter as well.

The last point about variance estimation is that the clusters in the population described in Section 4 need not be the sampling clusters (PSU's) of Section 5. We can *assume* they are, but that forces us to make an unverifiable assumption. Moreover, this assumption does not always render $\mathbf{V}_{RB}$ asymptotically unbiased when the sampling mechanism is not ignorable.

Both the extended and standard linear models were easily generalized to a class of non-linear models (see equation (9)). Pfeffermann et al. (1998) generalize the standard model to hierarchical structures. The applicability of the extended model in the hierarchical context is less clear. There are some assumptions about the structure that simply have to be made. Similarly, Kott (1996) provides a not-very-satisfying treatment of instrumental-variable regression under the extended model. Again, the problem is that additional assumptions have to be made that violate the spirit of the extended model. The standard model has great appeal when, due to random errors in the explanatory variables, one chooses to use instrumental-variable regression. Conducting an instrumental-variable regression with survey data is a different matter entirely from using an instrumental variable in calibration. The latter (only) is discussed in Kott (2003).

References

Asok, C., & Sukhatme, B. V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, *71*, 912-918.

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.

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

Korn, E. L., & Graubard, B. I. (1976). Variance estimation for superpopulation parameters. *Statistica Sinica*, *8*, 1131-1151.

Kott, P. S. (1991). A model-based look at linear regression with survey data. *The American Statistician*, *45*, 107-112.

Kott, P. S. (1996). Linear regression in the face of specification error: a model-based exploration of randomization-based techniques. *SSC Proceedings of the Survey Methods Section*, 39-47.

Kott, P. S. (2003). A practical use for instrumental-variable calibration. *Journal of Official Statistics*, *19*, 265-272.

Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society*, *60*(Series B), 115-126.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, *60*(Series B), 23-40.

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.

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

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