REGRESSION ANALYSIS OF REPEATED SURVEY DATA (WITH AVAILABLE SOFTWARE)

Phillip S. Kott, National Agricultural Statistics Service
NASS, 3251 Old Lee Highway, Room 300, Fairfax, VA 22030

Key Words: Design based; Enumeration unit, Extended linear model; Infinite population regression coefficient; Primary sampling unit

1. Introduction

A scientist usually thinks of linear regression as a means of estimating the parameters of a preconceived linear model or of testing the validity of a particular model within a continuum of slightly more general linear models. According to this "model based" theory of linear regression, part of the multivariate data - the dependent variable - is itself a random variable generated by a stochastic model.

In contrast, most survey statisticians favor an orthodox "design based" theory in which all the data are fixed values; the only thing probabilistic is the selection process that randomly chooses some data points for the sample and not others. There is no model generating the data. There is only a useful way of summarizing the covariation of multivariate values in the finite population: ordinary least squares applied to the entire population.

Orthodox design based theory may be mathematically appealing but it is scientifically sterile. This approach to inference can tell us nothing about the processes that shape the world since its only concern is correctly describing fixed, finite populations.

Fortunately, there is an alternative school of thought in design based theory, which we will call "infinite population design based." This approach to inference, advocated by Fuller (1975 & 1984), holds that there *is* an underlying model generating the data, but that the analyst knows very little about it. In fact, the relationship among the variables may not even be linear. Linear regression is simply a means of summarizing in linear fashion a relationship among the multivariate values generated by the model. Surprisingly, the infinite population design based approach to the analysis of survey data receives no mention in the otherwise excellent collection of papers, Skinner et al. (1989).

Shah et al. (1977) discuss standard design based techniques for estimating regression coefficients and their variance given a stratified, multi-stage sampling design incorporating with replacement sampling in the first stage of selection. The same techniques are recommended by Fuller (1975) for infinite population design based inference when the first stage of sampling using is conducted without replacement. Kott (1991a) shows that these design based techniques can also have useful properties from a model based perspective when there are missing regressors in the model and/or the error variance matrix is only vaguely specified.

Several software packages perform linear regressions and estimate variances using the design based techniques discussed in Shah et al. Two popular ones are SUDAAN (Shah et al. (1991)) and PC CARP (Fuller et al. (1986)).

This paper considers the application of linear regression to data from a survey repeated over time from an infinite population design based and an extended model viewpoint. For inferences under the extended model, little changes as long as elements can not move across primary sampling units (original sampling clusters). For infinite population design based inferences, variance estimation is affected when the stratification changes over across survey periods, as we shall see.

Before proceeding, it will helpful to describe some examples of repeated surveys. Although the three example discussed below are all yearly surveys, the methodologies to be presented in this paper apply equally well to repeated survey employing other periodic schemes and to surveys that are repeated only once or at irregular intervals.

The Annual Survey of Manufactures of the US Census Bureau enumerates a fixed panel of economic establishments for five survey years. Establishments are selected with probabilities proportionate to size using Poisson sampling.

The June Enumerative Survey of the National Agricultural Statistics Service is a yearly survey of agricultural activity. Area segments are selected using a form of stratified simple random sampling, and all farms within those segments are enumerated. Every year, 20% of the segments are removed from last year's sample and replaced with an equal number of new randomly selected segments.

The Farm Costs and Returns Survey, also of the National Agricultural Statistics Service, enumerates a stratified simple random sample of farms each year. The selection process is independent across years and even the stratification can change. Nevertheless, some large farms can find themselves enumerated in more than one survey year.

One final introductory note. For simplicity, the issue of survey nonresponse is completely ignored in this paper.

2. The Conventional Linear Model

The conventional linear model assumes that the multivariate values of a population of M elements (observations) can be fit by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \qquad (1)$$

where $\mathbf{y} = (y_1, ..., y_M)'$, is an M x 1 vector of population values for a dependent variable;
$\mathbf{X}$ is an M x K matrix of population values for K independent variables or regressors;
$\boldsymbol{\beta}$ is a K x 1 vector of regression coefficients; and
ϵ is an M x 1 vector of disturbances or errors satisfying $E(\epsilon) = \mathbf{0}$ and $Var(\epsilon) = E(\epsilon\epsilon') = \sigma^2\mathbf{I}_M$.

If one knows $\mathbf{y}$ and $\mathbf{X}$, then the best linear unbiased estimator of $\boldsymbol{\beta}$ would be the ordinary least squares (OLS) estimator

$$\mathbf{B} = (\mathbf{X'X})^{-1}(\mathbf{X'y}). \qquad (2)$$

When data comes from a survey sample, however, $\mathbf{y}$ and $\mathbf{X}$-values are only known for a sample of m elements which has been selected at random in a manner that is assumed to be independent of ϵ.

The best linear unbiased estimator of $\boldsymbol{\beta}$ under the model given the information available is

$$\mathbf{b}_{OLS} = (\mathbf{X'SX})^{-1}(\mathbf{X'Sy}),$$

where S is an M x M diagonal matrix of 0's and 1's. The i'th diagonal of S is 1 if and only if the ith element of the population is in the sample.

The variance of $\mathbf{b}_{OLS}$ (a variance-covariance matrix) is $\sigma^2(\mathbf{X'SX})^{-1}$. Since $(\mathbf{X'SX})^{-1}$ is known, an unbiased estimator for this variance can be determined by estimating σ^2 with

$$s^2 = (\mathbf{y} - \mathbf{X}\mathbf{b}_{OLS})'\mathbf{S}(\mathbf{y} - \mathbf{X}\mathbf{b}_{OLS})/(m - K).$$

3. The Sample-Weighted Estimator

Let $\mathbf{P}$ be a M x M diagonal matrix whose ith diagonal is the probability unit i was selected for the sample. We can call $\mathbf{W} = (m/M)\mathbf{SP}^{-1}$ the matrix of sampling weights. Note that $\mathbf{W} = \mathbf{S}$ when every element has a probability of selection equal to m/M.

For many sampling designs the weighted regression estimator,

$$\mathbf{b}_W = (\mathbf{X'WX})^{-1}(\mathbf{X'Wy}), \qquad (3)$$

is a design consistent estimator of $\mathbf{B}$ in equation (2); that is, as m grows arbitrarily large, $\text{plim}_{m\to\infty}(\mathbf{b}_W - \mathbf{B}) = 0$ with respect to the probability space generated by the sampling mechanism.

Not only is $\mathbf{b}_W$ often a consistent estimator of the finite population regression coefficient $\mathbf{B}$, it is also often a consistent estimator the *infinite population regression coefficient* $\mathbf{B}^* = \mathbf{Q}^{-1}\mathbf{R}$, where $\mathbf{Q} = \lim_{M\to\infty}(\mathbf{X'X})/M$ and $\mathbf{R} = \lim_{M\to\infty}(\mathbf{X'y})/M$. All that is necessary is for $\mathbf{Q}^{-1}$ and $\mathbf{R}$ to exist and $\mathbf{b}_W$ to be a consistent estimator of $\mathbf{B}$. See Fuller (1975) for details.

Unlike orthodox design based theory, the infinite population design based approach to linear regression assumes the existence of a model generating the finite population data. It does not assume very much about the nature of that model, however. This approach employs the laws of probability in the same way as the orthodox design based theory does: through the sample selection process exclusively.

Kott (1991a) observes that $\mathbf{b}_W$ can also be justified from a purely model based perspective by extending the linear model in equation (1) and assuming that the multivariate values of the population of M elements can be fit by the linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z} + \epsilon, \qquad (4)$$

where $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and ϵ are as before except that $Var(\epsilon)$ need not equal $\sigma^2\mathbf{I}_M$. The new vector $\mathbf{z}$, the *putative missing regressor*, satisfies $\lim_{M\to\infty}\mathbf{X'z}/M = 0$. It is a composite of all the regressors in a fully specified model for $\mathbf{y}$ that are otherwise missing from equation (1) and whose joint effect on $\mathbf{y}$ can not be captured within $\mathbf{X}\boldsymbol{\beta}$.

Under mild conditions, $\mathbf{b}_W$ is nearly (i.e., asymptotically) unbiased under the model (as n grows large). The same can not be said for $\mathbf{b}_{OLS}$ unless $\lim_{M\rightarrow\infty}\mathbf{X'Pz}/m = 0$, which in practical terms means that the probabilities of selection are unrelated to the missing regressors. Proofs of these assertions are in Kott (1991a).

4. Adding a Time Dimension

Suppose now that the sample does not come from a one time survey but from T surveys of the same population of enumeration units (individuals, households, firms, or whatever). This population may be dynamic; that is, enumeration units may enter or leave the population across survey periods.

At each survey period t, the population consists of M_t elements, while the sample consists of m_t elements. The joint population across all T periods consists of $M = \Sigma^T M_t$ elements, while the joint sample contains $m = \Sigma^T m_t$ elements. This means that the same enumeration unit (e.g., a farm in the June Enumerative Survey) is considered a different element in every survey period in which it is part of the population.

Equations (1) and (4) look the same as before. The difference is that the M-vector **y** can may now contain multiple values for the same enumeration unit, one from every survey period the unit is in the sample. A single model parameter, $\boldsymbol{\beta}$, appears to apply in every survey period. Appearances are misleading, however. To see why, consider the following example, which is trivial mathematically but often useful in practice. Let the i'th row of **X** be $(\delta_{i1}, \delta_{i2}, ..., \delta_{iT})$, where $\delta_{it} = 1$ when i=t and 0 otherwise. Now $\boldsymbol{\beta} = (\beta_1, ..., \beta_T)'$ also have T members. In fact, β_t is the mean of the y-values for survey period t under the stipulated model. Note that each β_t is associated with an distinct survey period.

Returning to the more general case, the framework for a defining the infinite population regression coefficient, $\mathbf{B}^*$, has to be clarified. We will assume here that the number of survey periods is fixed. As the (joint) population grows arbitrarily large, M_t/M_s remains constant for all $t \neq s$. We will also assume for asymptotic (i.e., large *sample*) analyses that as the sample grows arbitrarily large, all m_t/m_s remain constant.

Addressing the issue of variance (or mean squared error) estimation from the infinite population design based or extended model viewpoint requires a greater degree of specificity about the T sampling designs than has yet been provided. In what follows, we assume that the enumeration units selected for each survey period are chosen using a stratified, multistage probability sampling design. Unstratified and single stage surveys are special cases of this general sampling framework. Primary sampling units (PSU's) can be selected using either equal or unequal probability sampling, without replacement.

In the next section, we will further assume that the same n PSU's are randomly selected at the first stage of sampling for every survey period. We will, however, allow the enumeration units subsampled from these PSU's to change across periods.

There is an obvious asymmetry in the way elements and PSU's have been treated. Unlike elements, PSU's do not change identities from one survey period to the next. With a single stage survey design, like that used for the Annual Survey of Manufactures, the PSU's are the enumeration units. Consequently, there may be as many of T elements (observations) associated with a single PSU.

An additional assumption we will make in the next section is that each enumeration unit must remain in the same PSU across survey periods. It may enter or leave the population, but it can not change PSU's.

5. Variance Estimation -- The Restricted Case

Suppose that for sampling purposes the population has been divided into H strata (H may equal 1). Suppose further that there are at least two randomly selected PSU's from each stratum h.

Let n_h be the number of PSU's selected from stratum h. We can rewrite $\mathbf{b}_W$ in equation (3) as

$$\mathbf{b}_W = \sum_{h=1}^{H} \sum_{j=1}^{n_h} (\mathbf{X'WX})^{-1}\mathbf{X'WD}_{hj}\mathbf{y},$$

where $\mathbf{D}_{hj}$ as a M x M diagonal matrix of 1's and 0's such that the ith diagonal of $\mathbf{D}_{hj}$ is 1 if and only if the ith member of **y** corresponds to an element in PSU hj.

We assume here that in the definition of the infinite population regression coefficient, $\mathbf{B}^*$, *the number of PSU's in the population of each stratum tends to infinity in proportion to M.* Korn and Graubard (1994) explore an alternative framework where the relative stratum population sizes are random.

In many economic surveys, some PSU's are selected with certainty. In the infinite population design based framework adopted here, such certainty selections would be grouped into strata and treated as if they were randomly selected draws within strata.

We also assume here that the sampling design is such that the $\mathbf{f}_{hj} = \mathbf{Q}^{-1}\mathbf{X}'\mathbf{W}\mathbf{D}_{hj}(\mathbf{y} - \mathbf{X}\mathbf{B}^*)$ for $j = 1, ..., n_h$ are identically distributed and independent from the infinite population design based viewpoint. This is equivalent to treating the sample of PSU's within a stratum as if it had been selected from an infinite population *with replacement.* The approach taken here can effectively change some of the properties of many probability proportional to size sampling designs. The framework adopted by Korn and Graubard (1994) may be superior in that regard. The appendix sheds some light on this issue.

Under the assumptions adopted here, $\mathbf{f}_{hj} \approx m\gamma_{hj}$ when n is large, where $\gamma_{hj} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{D}_{hj}$ $(\mathbf{y} - \mathbf{X}\mathbf{B}^*)$. As a result, $\mathbf{b}_W - \mathbf{B}^* = \Sigma^H \Sigma^n \gamma_{hj}$, and

$$\mathbf{v}_\gamma = \sum_{h=1}^{H} n_h(n_h-1)^{-1}[\sum_{j=1}^{n_h} \gamma_{hj}\gamma_{hj}' - n_h^{-1}(\sum_{j=1}^{n_h} \gamma_{hj})(\sum_{j=1}^{n_h} \gamma_{hj})']$$

is a nearly unbiased estimator for the variance (actually the mean squared error matrix) of $\mathbf{b}_W$ from the infinite population design based viewpoint.

Unfortunately, the γ_{hj} are unknown, so $\mathbf{v}_\gamma$ can not be calculated in practice. Since $\mathbf{g}_{hj} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ $\mathbf{X}'\mathbf{W}\mathbf{D}_{hj}(\mathbf{y} - \mathbf{X}\mathbf{b}_W) \approx \gamma_{hj}$, the following practical estimator for the infinite population design based variance of $\mathbf{b}_W$ immediately presents itself:

$$\mathbf{v} = \sum_{h=1}^{H} n_h(n_h-1)^{-1}[\sum_{j=1}^{n_h} \mathbf{g}_{hj}\mathbf{g}_{hj}' - n_h^{-1}(\sum_{j=1}^{n_h} \mathbf{g}_{hj})(\sum_{j=1}^{n_h} \mathbf{g}_{hj})'] \qquad (5)$$

The variance estimator in equation (5) is computed by the SUDAAN software package when the design is specified as being with replacement in the first stage. PC CARP scales $\mathbf{v}$ by $\{(m-1)/(m-K)\}$.

One thing should be kept in mind when using these design based software packages for analyzing repeated survey data: the data must to sorted by stratum and PSU. That is to say, elements from different periods but the same PSU must be grouped together.

Let us return to the extended model in equation (4) and assume that $\mathbf{z} \equiv \mathbf{0}$ so that $\mathbf{b}_W$ is an model unbiased estimator of β. Following the logic in Kott (1991a), $\mathbf{v}$ in equation (6) is an nearly unbiased estimator for the variance of $\mathbf{b}_W$ under mild conditions so long as $E(\varepsilon_i\varepsilon_k)$ is zero when i and k are elements from different PSU's and bounded otherwise. A more efficient variance estimator (i.e., one with more "degrees of freedom") is

$$\mathbf{v}' = n(n-1)^{-1} \sum_{h=1}^{H} \sum_{j=1}^{n_h} \mathbf{g}_{hj}\mathbf{g}_{hj}', \qquad (6)$$

which equals $\mathbf{v}$ when $H = 1$ and $\Sigma\, \Sigma\, \mathbf{g}_{hj} \equiv \mathbf{0}$. Both $\mathbf{v}$ and $\mathbf{v}'$ rely on the fact that the $m(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{D}_{hj}(\mathbf{y} - \mathbf{X}\beta) = m(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{D}_{hj}\epsilon$ $\approx m\mathbf{g}_{hj}$ $(j = 1, ..., n_h)$ are independent random variables under the model with mean $\mathbf{0}$.

6. Variance Estimation -- The Less Restricted Case

Few repeated survey data sets come from surveys with the restrictive design assumed in the last section. In this section we generalize the results of the last section by allowing the first stage sample of PSU's to vary from survey period to survey period, as it does in the June Enumerative Survey. In fact, the stratification itself may change across periods as in does in the Farm Costs and Returns Survey. Moreover, PSU's can either come into existence or leave the population over time. As before, however, enumeration units can not change PSU's.

Let n be the number of PSU's selected (in the first stage of sampling) for at least one survey period, and let H_t be the number of strata in survey period t. Each of the n PSU's represented in the survey data set is in one of the H_t strata for survey period t or is not in the sample for that period. We can classify these n PSU's into H variance strata based on their stratification in each period. PSU's j and k are classified in the same variance strata when they are in the same stratum in every period they are sampled; when one is not sampled, neither is the other. Observe that $H \leq \Sigma^T H_t + T$.

As an example of how variance strata are formed, consider a data set consisting of two consecutive years of June Enumerative Survey data. Recall that 20% of the sample is rotated out of the survey every year and replaced by incoming area segments (PSU's). Let us focus on an element in the data set from a particular design stratum containing 10 sampled area segments in each survey year. This element would be allocated to one of three variance strata: a variance stratum containing elements from the two area segments sampled in the first survey year but not the second, a variance stratum containing elements from the two area segments sampled in the second survey year but not the first, or a variance stratum containing elements from the eight area segments sampled in both years.

From the infinite population design based viewpoint taken here, equation (5) still provides a nearly unbiased mean squared error estimator for $\mathbf{b}_W$ under certain conditions as long as n_h, the number of PSU's in variance stratum h with elements in the data set, is never equal to 1. The reasoning is unchanged from last section: if the sampling design is such that the $\mathbf{mg}_{hj}$ ($j = 1, ..., n_h$) are nearly independent and identically distributed, then equation (5) provides a nearly unbiased estimator for the infinite population design based variance of $\mathbf{b}_W$.

When some n_h is equal to 1, variance strata will have to be collapsed for variance estimation purposes. This will, if anything, lead to variance estimates with a slight upward bias (see Wolter 1985). In practice, there may have to be a good deal of collapsing. The resulting bias, however, may be small, as we shall see.

From the viewpoint of the extended model (with $\mathbf{z} \equiv \mathbf{0}$), equation (5) again provides a nearly unbiased estimator for the variance of $\mathbf{b}_W$, but equation (6) is better. The reasoning is the same as before: the $m(\mathbf{X'WX})^{-1}\mathbf{X'WD}_{hj}\epsilon$ are independent random variables with mean zero. Note that equation (6) provides the extreme case of collapsing variance strata when $\sum \sum \mathbf{g}_{hj} = \mathbf{0}$.

7. Contrasting Aspects of the Two Approaches to Inference

We have seen how well-known design based software packages can be used to perform linear regressions on repeated survey data in a scientifically meaningful manner. What is principally required for variance estimation is that PSU's and variance strata be defined appropriately for the package at hand. In particular, the variance strata described in the last section should be treated as strata, while the "sample size" of PSU's per variance strata is the number of distinct PSU's across all survey periods.

The incorporation of sampling weights in $\mathbf{b}_W$ is justified from the extended model viewpoint because it allows for the possibility that the conventional linear model in equation (1), which does not contain a $\mathbf{z}$ term, is misspecified. On the other hand, $\mathbf{z}$ is treated as if it were zero in variance estimation under the extended model. This logical inconsistency has a practical explanation.

In many applications, the elements of $\mathbf{z}$ tend to be very small in absolute value compared to those of the random error vector, ϵ. In fact, they are so small that their contribution to the mean squared error of $\mathbf{b}_W$ can be ignored. If we fail to incorporate the weights in $\mathbf{b}_W$, however, there may be a bias in the estimator caused by $\mathbf{z}$ being non-zero. This bias does not decrease as the sample size increases. By contrast, the variance of $\mathbf{b}_W$, which is mainly the result of ϵ being non-zero, does decrease with the sample size. Consequently, it may be prudent to remove the effect of a small, non-zero $\mathbf{z}$ when estimating $\mathbf{b}_W$ even though that same $\mathbf{z}$ has an ignorable impact on the variance of $\mathbf{b}_W$.

The assumption of the extended model that dominates variance estimation is that $E(\varepsilon_i\varepsilon_k) = 0$ when i and k are from distinct PSU's and bounded otherwise. There are two reasons why a PSU may contain more than a single element. Sampling designs can be clustered or the same enumeration unit can be selected more than once across survey periods. For certain surveys and regression equations, it may be reasonable to assume that while observations from the same enumeration unit are correlated, observations from different enumeration units in the same sampling cluster are not (conditioned on their respective $\mathbf{x}_i$ vectors). If this assumption is true, then it makes sense from a model based perspective to redefine the enumeration units as the "PSU's" when using a design based regression package. This increases the efficiency of the resulting variance estimator.

The justification for the variance estimators in equations (5) and (6) from the extended model viewpoint relied on the $\mathbf{mg}_{hj} = (\mathbf{X'WX})^{-1}\mathbf{X'WD}_{hj}$ $(\mathbf{y} - \mathbf{Xb}_W)$ having a mean of $\mathbf{0}$. By contrast, in the infinite population design based approach to inference taken here, the $\mathbf{f}_{hj} = \mathbf{Q}^{-1}\mathbf{X'WD}_{hj}(\mathbf{y} - \mathbf{XB}^*)$ $\approx \mathbf{mg}_{hj}$ may be identically distributed within strata, but they did not have means of $\mathbf{0}$. Heuristically, the latter approach allows for the putative missing regressor to have different impacts on different strata.

In practice, if the use of equation (5) leads to appreciably smaller estimates of the variance of $\mathbf{b}_W$ across the diagonals than does equation (6), then the model in equation (1) may be underspecified. A potential remedy is to add regressors to the model that are related to the stratification, such as separate dummies for each stratum or variance stratum.

8. An Example

In this section, we explore a synthetic example that may shed some light on a number of issues. Consider a population covering two survey periods. The population consists of 50 business entities (enumeration units) in Period 1. By Period 2, eight of these entities have gone out of business, and two new entities has emerged.

As far as the sampling design goes: in Period 1, ten entities were selected for the sample with certainty, while another ten were selected via simple random sampling without replacement (srswor). In Period 2, eight of the original ten certainties were again selected with certainty, while an additional twelve entities were selected via srswor independently of the selections in Period 1. None of the period 1 certainties went out of business, and neither of the two emerging entities were Period 2 certainties.

There are 94 elements in the joint population (50 + 44) and 52 PSU's (entities). Let us label the entities in Period 1, 1 through 50, and the two new entities in Period 2, 51 and 52. Suppose the sample from each design stratum looked like this:

$S_{1C} = \{1, 2, ..., 10\}$
$S_{1P} = \{11, 12, ..., 20\}$
$S_{2C} = \{1, 2, ..., 8\}$
$S_{2P} = \{10, 11, 12, 21, 22, ..., 28, 51\}$,

where the first subscript of S_{xz} denotes the period (1 or 2) and the second whether the stratum contains certainty (C) or probability (P) selections.

Most of the sampled entities can be grouped into one of four variance strata:

$V_1 = \{1, 2, ..., 8\}$ consists of certainties for both periods,
$V_2 = \{11, 12\}$ consists of random selections for both periods,
$V_3 = \{13, 14, ..., 20\}$ consists of random selections for Period 1 only, and
$V_4 = \{21, ..., 28, 51\}$ consists of random selections for Period 2 only.

This leaves entities 9 and 10, each of which merits its own distinct variance stratum. Unfortunately, there must be at least two PSU's in each variance stratum, so some collapsing is required. One simple solution is to create
$V_5 = \{9, 10\}$.

Observe that the fact that Entity 51 is new for Period 2 plays no part in variance stratum assignment. Similarly, we are not concerned about entities in V_3 that may have gone out of business after Period 1.

Let y_{jt} be a value of interest for entity j in Period t. Each jt denotes a distinct element in the population.

One very simple use of linear regression is to estimate the change in average y values between Periods 1 and 2. This can be done, for example, with the following regression equation:

$$y_{jt} = \beta_0 + \beta_1 x_{jt} + e_{jt}, \qquad (7)$$

where $x_{jt} = 1$ when $t = 2$, and zero otherwise.

If OLS were performed on the entire population, then one would find that the finite population regression coefficient $B_1 = \Sigma_{U2}\ y_{j2}/44 - \Sigma_{U1}\ y_{j1}/50$, where Σ_{Ut} means summation over the population at Period t. The sample weighted estimator for B_1 is $b_{1W} = \Sigma_{S2}\ w_{j2}y_{j2}/\Sigma_{S2}\ w_{j2} - \Sigma_{S1}\ w_{j1}y_{k1}/\Sigma_{S2}\ w_{j1}$, where St denotes the sample for period t, and $w_{jt} = 1$ if entity j is a certainty selection for period t, 4 if $j > 10$ and $t = 1$, or 3 if $j > 8$ and $t = 2$.

The finite population coefficient B_1 measures the difference between the average y value among the 44 entities in Period 2 and the 50 entities in Period 1, but sometimes we are interested in more. We may want to draw inference about different periods in general or, less grandly, about the conditions that caused the entities in the two periods under examination to be as they are. For that, we can estimate the infinite population regression coefficient: $B_1^* = y_2^* - y_1^*$, where $y_t^* = \lim_{N\to\infty}\{\Sigma_{Ut}\ y_{jt}/N_t\}$, N_t is the number of entities in the population at period t, and $N = N_1 + N_2$. Happily, its estimator is also b_{1W}.

Let $g_{jt} = w_{jt}(y_{jt} - \Sigma_{St}\ w_{kt}y_{kt}/\Sigma_{St}\ w_{kt})/\Sigma_{St}\ w_{kt}$. The variance estimator for b_{1W} as an estimator for B_1^* has the form of equation (5) with five variance strata (V_1,..., V_5). Observe that $n_1 = 8$, $n_2 = 2$, $n_3 = 8$, $n_4 = 9$, $n_5 = 2$, and $n =29$. Each $\mathbf{g}_{hj}$ in equation (5) is represented by $g_{j+} = d_{j2} - d_{j1}$, where $d_{jt} = g_{jt}$ when $j \in St$ and 0 otherwise.

The key to variance estimation from the infinite population design based viewpoint taken here is that the ng_{j+} are nearly independent and, except in the collapsed stratum V_5, nearly identically distributed. This is because each ng_{jt} is approximately equal to $q_{jt} = \lim_{N\to\infty} nw_{jt}(y_{jt} - y_t^*)/N_t$, and the q_{jt} within each of the original design strata (V_{1C}, V_{1P}, V_{2C}, and V_{2P}) are identically distributed. As a result, the $q_{j+} = q_{j2} - q_{j1}$, where $\delta_{jt} = q_{jt}$ when $j \in St$ and 0 otherwise, are identically distributed within V1 through V4, respectively; not surprising, since the variance strata were created for this very purpose! All the q_{j+} are PSU-specific and so assumed to be independent from the infinite population design based viewpoint taken here, even those in V_5. The fact that the q_{j+} in V5 are not identically distributed may bias the variance estimator upward.

From the extended model viewpoint, the goal is to estimate the model parameter β_1 in equation (7). Let $e_{jt} = z_{jt} + \varepsilon_{jt}$, where ε_{jt} is a random variable with mean zero, and z_{jt} is the putative missing regressor. If $z_{jt} \equiv 0$, or if its size is very small compared to the error term ε_{jt}, then all the PSU-specific ng_{j+} are nearly independent and have approximately the same model expectation, 0 (since each $ng_{jt} \approx nw_{jt}\varepsilon_{jt}/\Sigma_{St}\, w_{kt}$). As a result, the variance estimator discussed above is also an estimator of the model variance of b_{1W} as an estimator for β_1. A more efficient variance estimator would have the form of equation (6) with g_{j+} again replacing $\mathbf{g}_{hj}$.

In this particular example, equation (7) has only a single explanatory variable. Consequently, it may not be very reasonable to suppose that the putative missing regressor has no influence of the variance of b_{1W}. Infinite population design based inference seems superior to inference under the extended model for this example.

Kott (1991b) discusses the estimation of a system of linear equations. In this context, the analyst has a strong belief in the completeness the model, and it is infinite population design based inference that fails to be a useful statistical tool.

9. Two Final Comments

9.1 The Jackknife

Equation (5) takes the form of the so-called "linearization variance estimator" computed by SUDAAN and PC CARP when one stipulates with replacement sampling in the first stage of selection. A well-known alternative to this estimator is the jackknife; see, for example, Krewski and Rao (1981, equation (2.4)). The jackknife variance estimator for $\mathbf{b}_W$ conformal to equation (5) is

$$v_J = \sum_{h=1}^{H} (n_h - 1)/n_h \sum_{j=1}^{n_h} (\mathbf{b}_{W(hj)} - \mathbf{b}_W)(\mathbf{b}_{W(hj)} - \mathbf{b}_W)', \tag{8}$$

where $\mathbf{b}_{W(hj)} = (\mathbf{X'W}_{(hj)}\mathbf{X})^{-1}\mathbf{X'W}_{(hj)}\mathbf{y}$, and

$$\mathbf{W}_{(hj)} = \mathbf{W}(\mathbf{I}_M + 1/(n_h-1) \sum_{j=1}^{n_h} \mathbf{D}_{hg} - (n_h/[n_h-1])\mathbf{D}_{hj}).$$

Observe that

$$\begin{aligned}\mathbf{b}_{W(hj)} - \mathbf{b}_W &\approx (\mathbf{X'WX})^{-1}\mathbf{X'W} \\ &\quad [1/(n_h-1)\Sigma^n\, \mathbf{D}_{hg} - (n_h/[n_h-1])\mathbf{D}_{hj}](\mathbf{y} - \mathbf{XB}^*) \\ &\approx (\mathbf{X'WX})^{-1}\mathbf{X'W} \\ &\quad [1/(n_h-1)\Sigma^n\, \mathbf{D}_{hg} - (n_h/[n_h-1])\mathbf{D}_{hj}](\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &\approx -\,\mathbf{g}_{hj} + \sum_{g\neq j}^{n_h} \mathbf{g}_{hg}/(n_h-1)\end{aligned}$$

under mild conditions. It is now not hard to see that given the same definitions of PSU's, variance strata, and PSU sample sizes, the Jackknife variance estimator in equation (8) is nearly unbiased in both the extended model and infinite population design based senses whenever the linearization variance estimator in equation (5) is. Rust (1985) discusses a computational simplified jackknife where the PSU's within a stratum are randomly grouped, and then a group is deleted one at a time. In other words, the groups become the variance PSU's in equation (8). This simplification is a practical necessity for many single stage surveys where the number of PSU's can be in the thousands.

If the original PSU's and variance strata are defined as in the text, and n_h is the number of variance PSU's in variance stratum h, then it is not hard to show that Rust's computationally simplified jackknife produces a variance estimator that is also nearly unbiased in the extended model sense whenever the analogous linearization form is. From the infinite population design based viewpoint taken here, each variance PSU in a variance stratum much contain the same number of original PSU's (so that the $\mathbf{b}_{W(hj)}$ have identical asymptotic distributions for each variance PSU j in variance stratum h).

9.2 A Warning About Enumeration Units Changing PSU's

When developing variance estimators in the text, we assumed that enumeration units can not change PSU's from one survey period to another. This will be the case for many repeated surveys, especially economic surveys where samples are drawn from list frames.

Most demographic surveys, however, are based on area frames. When the enumeration unit for a repeated demographic survey is a household or an individual, it is not uncommon for some enumeration units to relocate from one PSU to another across survey periods. Thus, the following point needs to be underlined: When two sampled elements in the data set from a repeated survey are associated with the same enumeration unit but different PSU's, the variance estimators discussed in this paper do not formally apply. It is easy to see from the extended model viewpoint that the independence of sampled elements from different PSU's is lost when this happens.

References

Asok, C. and Sukhatme, B.V. (1976), "On Samford's of Unequal Probability Sampling Without Replacement," *Journal of the American Statistical Association*, 71, 912-918.

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhya*, Ser. C, 37, 117-132.

____ (1984), "Least Squares and Related Analyses for Complex Survey Designs," *Survey Methodology*, 10, 97-118.

Fuller, W. A., Kennedy, W., Schnell, D., Sullivan, G., and Park, H. J. (1986), *PC CARP*, Ames, IA: Statistical Laboratory, Iowa State University.

Kott, P. S. (1991a), "A Model-Based Look at Linear Regression with Survey Data, *American Statistician*, 107-112.

____ (1991b), "Estimating a System of Linear Equations with Survey Data," *Survey Methodology*, 91-98.

Korn, E.L. and Graubard, B.I. (1994), "Variance Estimation for Superpopulation Parameters: Should One Use With Replacement Estimators?" *American Statistical Association, Proceedings of the Survey Research Methods Section*, this volume.

Krewski, D. and Rao, J.N.K. (1981), "Inferences from Stratified Samples: Properties of Linearization, Jackknife, and Balanced Repeated Replication Methods," *Annals of Statistics*, 9, 1010-1019.

Rust, Keith (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics*, 1, 381-397.

Shah, B. V., Holt, M. M., and Folsom, R. E. (1977), "Inference About Regression Models from Sample Survey Data," *Bulletin of the International Statistical Institute*, 47, 43-57.

Shah, B. V., Barnswell, B. G., Hunt, P. N., and LaVange, L. M. (1991), *SUDAAN™ User's Manual*, Release 5.50, Research Triangle Park, NC: Research Triangle Institute.

Skinner, C.J., Holt, D, and Smith T.M.F. (1989), *Analysis of Complex Surveys*, New York: Wiley.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, pp. 47-54, New York: Springer-Verlag.

The appendix is available from the author upon request.