HYPOTHESIS TESTING OF LINEAR REGRESSION COEFFICIENTS WITH SURVEY DATA

Phillip S. Kott, U.S. Bureau of the Census
U.S. Bureau of the Census, Room 3061-3, Washington, D.C. 20233

Key Words: Design-based; Model-dependent; Probability order; Effective degrees of freedom; Bonferroni.

1 INTRODUCTION

Most of statistical theory is analytical in nature. One begins with a set of data and a fairly general stochastic model believed to have generated that data. Statistical theory is then invoked to estimate the parameters of the model and to determine the accuracy of those estimates. Ultimately, the original model may be pared down as the result of a series of statistical tests which often take the form of investigations into whether particular parameter values may be reasonably inferred to be zero.

The bulk of survey sampling theory, by contrast, is not analytical but descriptive. There is a finite population of interest. Information about this population can, in principle, be summarized by means of one or more descriptive statistics (for example, the population mean and median). The survey statistician is constrained by time or budgetary considerations to estimate such statistics using only a sample of population units. He (she) often faces a two-fold problem: first a method of sample selection needs to be chosen, then the population statistic(s) needs to be estimated from the sample. Although it is possible to construct a model-based statistical theory for these purposes (see, for example, Royall, 1970), most survey statisticians invoke a model-free approach known as design-based sampling theory. In this theory, it is not the sample data values that are stochastic (as they are in model-based theory) but the sample selection process. Rao & Bellhouse (1989) provides a useful summary of both design and model-based theory and of attempts to synthesize the two approaches.

This paper is concerned with estimating parameters of a linear model based on data from a sample survey. Although design-based theory was originally intended for descriptive rather than analytical inference, Kish & Frankel (1974) and Fuller (1975) among others have generalized the results of that theory to the estimation of linear regression parameters. Theirs is <u>not</u> the tack taken here; rather, we take essentially a model-dependent approach to the problem. In so doing, however, we essentially build on design-based <u>techniques</u>. Both Skinner (1989) and Kott (1991) note the robustness of the design-based linearization variance estimator to complex error structures (i. e., the model variance matrix for the error term being block diagonal across primary sampling units but unspecified within blocks). The main concern here will be in testing hypotheses about linear

regression parameters. We will assume that the model is correct and that model errors are normally distributed with a possibly complex covariance structure. Unlike Wu, Holt, & Holmes (1988), however, we will not explicitly model the error structure (except, perhaps, at a very late stage). Rather, we will focus our attention on t-statistics calculated using the linearization variance estimator.

It will be argued that the design-based t-statistic for a particular regression coefficient should be modified by first reducing the bias of the linearization variance estimator and then estimating its effective degrees of freedom (which may require some error modelling). Since variance estimators for different regression coefficients are likely to have different effective degrees of freedom, a Bonferroni procedure is recommended for testing joint hypotheses about the coefficients.

2 A MOTIVATING EXAMPLE

The following example will demonstrate the need for adjusting conventional design-based practice in calculating and using t-statistics for regression coefficients.

Consider a simple random sample of n units, a of which are in a subset of the sample denoted by A. Let y_i be the observed value for unit i. Suppose the following linear model holds:

$$y_i = d_i b_1 + (1 - d_i) b_2 + \epsilon_i,$$

where $d_i = 1$ is unit i is in set A, and 0 if i is in A, the sample complement of A; and the ϵ_i are independent normally distributed random variables.

Assuming homoscedastic errors, both the model-based and design-based regression estimator for b_1 is the simple mean, $\bar{y}_A = \Sigma_{i\in A} y_i/a$. We will see that the linearization estimator for the variance of this estimator is

$$v_L = (n/[n-1])\Sigma_{i\in A}(y_i - \bar{y}_A)^2/a^2.$$

This differs from the model-based variance estimator:

$$v_M = \{\Sigma_{i\in A}(y_i - \bar{y}_A)^2 + \Sigma_{i\in \bar{A}}(y_i - \bar{y}_{\bar{A}})^2\}/[a(n-2)].$$

The advantage of v_L is that, unlike v_M, it is asymptotically unbiased under the model even if the ϵ_i are heteroscedastic. That point was noted by Skinner (1989) and Kott (1991). Unfortunately, there still may be considerable bias for finite n. For example, when n = 100 and a = 10, the relative bias of v_L is approximately 10%. We can see this by noting that

$$v_E = \Sigma_{i\in A}(y_i - \bar{y}_A)^2/(a[a-1]) = ([n-1]/n)(a/[a-1])v_L$$

is exactly unbiased.

If one were to calculate a t-statistic using conventional design-based practice, he (she) would not only use a biased variance estimator but would also assume that the statistic has 97 or 99 degrees of freedom (100 sampling units minus one strata minus two regressors, were this last subtraction is not always performed). Under ideal conditions (homoscedastic errors within set A), however,

the t-statistic calculated using v_E has a Student's t distribution with only 9 degrees of freedom.

3 THE MODEL

Suppose we have a population of M elements that can be fit by the linear model:

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

where $\mathbf{y}_M$ is an M x 1 vector of population values for the designated dependent variable; $\mathbf{X}_M$ is an M x K matrix of population values for the K designated independent variables; $\boldsymbol{\beta}$ is a K x 1 vector of regression coefficients; and $\boldsymbol{\epsilon}_M$ is a normally distributed random vector with mean $\mathbf{0}_M$ and variance Σ_M.

A random sample, S, of m distinct elements is drawn from the population. To allow a certain amount of generality in the sampling design, we assume that the population is divided into H strata. From each stratum h, n_h distinct clusters of elements are randomly sampled and denoted h1, h2, ..., hn_h. A random sample of m_{hj} elements is selected from each cluster hj. The clusters are also referee to as primary sampling units. There are $n = \Sigma\, n_h$ primary sampling units in the sample.

Each sampled element has a designation hji, where h is its stratum, hj its primary sampling unit within h, and i the element itself within hj. Let p_{hji} be the probability that element hji is in the sample, and let $w_{hji} = m/(Mp_{hji})$ be the sampling weight of the element.

The linear model in (1) also applies to the elements in sample S:

$$\mathbf{y}_S = \mathbf{X}_S\boldsymbol{\beta} + \boldsymbol{\epsilon}_S,$$

where $\mathbf{y}_S$, for example, is the m x 1 vector of sampled values for the dependent variable. Let $\boldsymbol{\epsilon}_{hj}=(\epsilon_{hj1}, \epsilon_{hj2}, \ldots, \epsilon_{hjm_{hj}})'$ be the error vector for the elements in primary sampling unit hj. Now, $\boldsymbol{\epsilon}_S$ can be arranged so that the $\boldsymbol{\epsilon}_{hj}$ are stacked one on top of the other. Let $\mathrm{Var}(\boldsymbol{\epsilon}_{hj}) = E(\boldsymbol{\epsilon}_{hj}\boldsymbol{\epsilon}_{hj}')$ be denoted by the m_{hj} x m_{hj} matrix Σ_{hj}, which need not be diagonal. We assume that the $\boldsymbol{\epsilon}_{hj}$ are uncorrelated across primary sampling units, so that Σ_S is block diagonal. The design-based estimator for $\boldsymbol{\beta}$ is the weighted least squares estimator:

$$\hat{\boldsymbol{\beta}}_W = (\mathbf{X}_S'\mathbf{W}\mathbf{X}_S)^{-1}\mathbf{X}_S'\mathbf{W}\mathbf{y}_S,$$

where $\mathbf{W}$ is the m x m diagonal matrix of sampling weights. The g^{th} diagonal value of $\mathbf{W}$ is the sampling weight associated with the g^{th} element of the sample. Clearly, $\hat{\boldsymbol{\beta}}_W$ is an unbiased estimator of $\boldsymbol{\beta}$ under the model in (1).

Kott (1991) shows that $\hat{\boldsymbol{\beta}}_W$ can remain nearly model unbiased when the model in equation (1) is misspecified. We are assuming here, however, that (1) is correct. Consequently, the ordinary least squares estimator,

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}_S'\mathbf{X}_S)^{-1}\mathbf{X}_S'\mathbf{y}_S,$$

is every bit as unbiased as $\hat{\boldsymbol{\beta}}_W$. Despite this fact, we will continue focusing our analysis on $\hat{\boldsymbol{\beta}}_W$, since whatever applies to $\hat{\boldsymbol{\beta}}_W$ also applies to $\hat{\boldsymbol{\beta}}_{OLS}$ (and to other weighted regression estimators) with a straight-

forward modification of **C** below.

One can simplify the notation for $\hat{\boldsymbol{\beta}}_w$ by letting **C** be the k x m matrix $(\mathbf{X}_S'\mathbf{W}\mathbf{X}_S)^{-1}\mathbf{X}_S'\mathbf{W}$, so that $\hat{\boldsymbol{\beta}}_w = \mathbf{C}\mathbf{y}_S$. Let $\mathbf{D}_{hj}$ be a m x m diagonal matrix with 1's corresponding to the sampled elements of hj and 0's elsewhere. Furthermore, let $\mathbf{C}_{hj} = \mathbf{C}\mathbf{D}_{hj}$. Finally, let $\mathbf{r}_S = \mathbf{y}_S - \mathbf{X}_S\hat{\boldsymbol{\beta}}_w$ be the vector of residuals.

The Taylor series or linearization estimator for the mean squared error of $\hat{\boldsymbol{\beta}}_w$ (Shah, Holt, and Folsom, 1977) is

$$\mathbf{mse}_L = \sum_{h=1}^{H} (n_h/[n_h-1]) \sum_{j=1}^{n_h} \mathbf{A}_{hj}\mathbf{r}_S\mathbf{r}_S'\mathbf{A}_{hj}' \qquad (2)$$

where $\mathbf{A}_{hj} = \mathbf{C}_{hj} - n_h^{-1}\Sigma\mathbf{C}_{hg}$, and the summation is over all the primary sampling units in stratum h. The terms "Taylor series" and "linearization" refer to the derivation of $\mathbf{mse}_L$ using design-based sampling theory. Kott (1991) shows that $\mathbf{mse}_L$ is a nearly unbiased estimator of the model variance of $\hat{\boldsymbol{\beta}}_w$ under reasonable conditions.

It should be noted that in their derivation of $\mathbf{mse}_L$, Shah, Holt, and Folsom assumed that the primary sampling units were chosen with replacement. Here, as in Kott (1991), we are assuming that the primary sampling units are distinct which suggests that they were selected without replacement. The reason for this discrepancy is that the assurance of independence among the selected primary sampling units within a stratum in design-based theory and model-based theory has almost opposite requirements. The discrepancy goes away, however, if we assume that the primary sampling units were chosen without replacement but that the goal of design-based regression theory is not to estimate a finite population regression parameter but the limit of that parameter as the population (and the number of primary sampling units per stratum) grows arbitrarily large. See Fuller (1975).

If the model in equation (1) holds and H > 1, then there is an alternative to $\mathbf{mse}_L$ that is also nearly unbiased. It has the same form as equation (2) except that all n sampled primary sampling units are treated as if they came from a single stratum (H = 1). Since the alternative can be expressed using equation (2), there is no need to treat it separately in the analysis that follows.

4 THE CONVENTIONAL t-STATISTIC

The estimator $\hat{\boldsymbol{\beta}}_w$ is a K-vector. In this section will be interested in the t-statistic used to test the univariate hypothesis that $\mathbf{q}\boldsymbol{\beta} = h$ for some K element row vector $\mathbf{q} = (q_1, q_2, \ldots, q_K)$. The most common example of such an hypothesis addresses whether a particular element of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, say β_r, is zero. In this example, all of the q_t would be zero except q_r which would be 1; h would also be zero.

If the model in (1) and the null hypothesis that $\mathbf{q}\boldsymbol{\beta} = h$ are true, then

$$x = (\mathbf{q}\hat{\boldsymbol{\beta}}_w - h)/\{\mathbf{q}\mathbf{Var}(\hat{\boldsymbol{\beta}}_w)\mathbf{q}'\}^{1/2}$$

would be normally distributed with mean 0 and variance 1. If $\mathbf{Var}(\hat{\boldsymbol{\beta}}_w)$ were known, then

the null hypothesis could be tested by comparing the statistic x to a standard normal table. Unfortunately, $\mathbf{Var}(\hat{\beta}_W)$ must be estimated from the sample. Conventional design-based practice is to compare the "t-statistic:"

$$t=(\mathbf{q}\hat{\beta}_W - h)/(\mathbf{q}\mathbf{mse}_L\mathbf{q}')^{1/2} \quad (3)$$

to a Student's t distribution with n - H or (n - H - K) degrees of freedom (see Shah, Holt, and Folsom, 1977).

The primary goal of this paper is to investigate and then modify the rather ad hoc practice described above using the model in equation (1) and our assumptions that Σ_S is block-diagonal. This will be done by examining the first four moments of t in (3). To this end, let

$$v^2 = \mathbf{q}\mathbf{Var}(\hat{\beta}_W)\mathbf{q}',$$
$$s^2 = \mathbf{q}\mathbf{mse}_L\mathbf{q}', \text{ and}$$
$$d = (s^2 - v^2)/v^2.$$

Under mild conditions, which we assume hold (see Appendix A), x is $O_P(1)$, and d is $O_P(n^{-1/2})$. Dropping terms of probability order $n^{-3/2}$, we have

$$\begin{aligned} t &\approx x(1 - d/2 + 3d^2/8), \\ t^2 &\approx x^2(1 - d + d2), \\ t^3 &\approx x^3(1 - 3d/2 + 15d^2/8), \\ \text{and } t^4 &\approx x^4(1 - 2d + 3d^2). \end{aligned} \quad (4)$$

In deriving (4), one assumes that n is large enough for $n^{-3/2}$ terms to be safely ignored. It is important to realize that this is weaker than a commonly made assumption that 1/n terms can be ignored. Under that stronger assumption, t would be approximately normal.

5 THE ADJUSTED t-STATISTIC

Appendix A shows that under mild conditions, $E(d) = O(1/n)$. Since we are not be dropping 1/n terms in this analysis, it seems advisable to reduce the slight model bias in s^2. This can be done by replacing s^2 with

$$s_*^2 = s^2/(1 - p), \quad (5)$$

where $p = s_0^{-2}\{\sum_{h=1}^{H}(n_h/[n_h - 1])\sum_{j=1}^{n_h}(\mathbf{g}_{hj} - \mathbf{g}_h)\hat{\mathbf{z}}(\mathbf{g}_{hj} - \mathbf{g}_h)'\}$,

$s_0^2 = \sum_{h=1}^{H}\sum_{j=1}^{n_h}\mathbf{g}_{hj}\hat{\Sigma}_S\mathbf{g}_{hj}'$,

$\mathbf{g}_{hj} = \mathbf{q}\mathbf{C}\mathbf{D}_{hj}$,

$\mathbf{g}_h = \sum_{j=1}^{n_h}\mathbf{g}_{hj}/n_h$,

$\hat{\mathbf{z}} = 2\mathbf{X}\mathbf{C}\hat{\Sigma}_S - \mathbf{X}\mathbf{C}\hat{\Sigma}_S\mathbf{C}'\mathbf{X}'$, and

$\hat{\Sigma}_S$ is block diagonal with each $\hat{\Sigma}_{hj} = \mathbf{r}_{hj}\mathbf{r}_{hj}'$.

Appendix B shows that under mild conditions, p is $O_P(1/n)$. This suggests that s_*^2 can have the same asymptotic design-based properties as s^2. From the model-based point of view taken here, $E(s_*^2/v^2) = 1 + O(n^{-3/2})$ under mild conditions. This is also demonstrated in Appendix B.

Let us define the adjusted t-statistic as

$$t_* = (\mathbf{q}\hat{\beta}_W - h)/s_*. \quad (6)$$

The analogue to equation (4) is

$$\begin{aligned} t_* &\approx x(1 - d_*/2 + 3d_*^2/8), \\ t_*^2 &\approx x^2(1 - d_* + d_*^2), \\ t_*^3 &\approx x^3(1 - 3d_*/2 + 15d_*^2/8), \\ \text{and } t_*^4 &\approx x^4(1 - 2d_* + 3d_*^2), \end{aligned} \quad (7)$$

where $d_* = (s_*^2 - v^2)/v^2$.

Let $e_{hj} = n\mathbf{g}_{hj}\epsilon_S$. The e_{hj} are

independent normally distributed random variables of probability order 1 under mild conditions. Let the variance of e_{hj} be denoted by v_{hj}^2.

Under the null hypothesis, x and d. can be re-expressed as (see Appendix C)

$$x = \Sigma\,\Sigma\, e_{hj}/(nv), \text{ and}$$
$$d_* \approx (nv)^{-2}\,\Sigma_h\{\Sigma_j[e_{hj}^2 - v_{hj}^2] - \Sigma_{i\neq j} e_{hj}e_{hi}/(n_h - 1)\}.$$

It is now not difficult to see that $E(t_*)$ and $E(t_*^3)$ are (approximately) 0. Appendix C establishes that

$$E(t_*^2) \approx 1 + 2(nv)^{-4}\sum_{h=1}^{H}\sum_{i\neq j} v_{hj}^2 v_{hi}^2/(n_h-1)^2, \quad (8)$$

and

$$E(t_*^4) \approx 3 + 2(nv)^{-4}\sum_{h=1}^{H}\{\sum_{j=1}^{n_h} 3v_{hj}^4 + [(6n_h + 3)/(n_h-1)^2]\sum_{i\neq j} v_{hj}^2 v_{hi}^2\}, \quad (9)$$

while the coefficient of kurtosis for t_* is

$$E(t_*^4)/[E(t_*^2)]^2 \approx 3(1 + 2[nv]^{-4}\sum_{h=1}^{H}\{\sum_{j=1}^{n_h} v_{hi}^4 + \sum_{i\neq j} v_{hj}^2 v_{hi}^2/[n_h-1]^2\}). \quad (10)$$

Let t_F be a random variable with a Student's t distribution with F degrees of freedom. If $F^{-3/2}$ is ignorable, then $E(t_F) = E(t_F^3) = 0$, $E(t_F^2) \approx (1 + 2/F)$, and $E(t_F^4)/[E(t_F^2)]^2 \approx 3(1 + 2/F)$. Comparing this last (near) equality to equation (10) suggests a Satterthwaite-like determination of the effective degrees of freedom of t_* (see Satterthwaite, 1946); namely,

$$F = \frac{(nv)^4}{\sum_{i=1}^{H}\{\sum_{j=1}^{n_h} v_{hj}^4 + \sum_{i\neq j} v_{hj}^2 v_{hi}^2/(n_h-1)^2\}} \quad (11)$$

which can be estimated from the sample by

$$f = \frac{(ns_0)^4 - \sum_{h=1}^{H}\sum_{j=1}^{n_h} 2s_{hj}^4/3}{\sum_{i=1}^{H}\{\sum_{j=1}^{n_h} s_{hj}^4/3 + \sum_{i\neq j} s_{hj}^2 s_{hi}^2/(n_h-1)^2\}} \quad (12)$$

where $s_{hj}^2 = n^2(g_{hj}r_S)^2$ (since $E(s_{hj}^2) = v_{hj}^2 + O(1/n)$, and $E(s_{hj}^4) = 3v_{hj}^4 + O(1/n)$; see also Appendix C).

The claim that t_* has approximately a Student's t distribution with F (or f) degrees of freedom under the null hypothesis relies on the twin assumptions that t_* is well described by its first four moments and that terms of order $n^{-3/2}$ are ignorable. Under those conditions, the claim may be slightly conservative because $E(t_F^2)$ is somewhat larger than $E(t_*^2)$ when the v_{hj}^2 are not all equal for a given h.

What is being recommended here is that one test whether $\mathbf{q}\beta = h$ by assuming under the null hypothesis that t_* in equation (6) has a Student's t distribution with either F or f degrees of freedom, where F is determined using equation (11) and making some assump-

tions about the v_{hj}, and f is determined using equation (12) and the data alone. Let us call this test the adjusted t-test. The advantage of using f over F is that the former requires less assumptions. On the negative side, f is a random variable and may be unstable.

6 JOINT HYPOTHESES

Often a statistician needs to test a hypothesis concerning more than a single regression coefficient. Many of these hypotheses can be put into $\mathbf{q}\beta = h$ form (for example, a hypothesis about the equality of two coefficients), but most cannot (for example, a hypothesis about the equality of three or more coefficients).

Conventional practice allows testing of joint hypotheses of the form $\mathbf{Q}\beta = \mathbf{h}$, where $\mathbf{Q}$ is an R x K matrix. The test statisistic is

$$T^2 = (\mathbf{Q}\hat{\beta}_W - \mathbf{h})'(\mathbf{Q}\mathbf{mse}_L\mathbf{Q}')^{-1}(\mathbf{Q}\hat{\beta}_W - \mathbf{h})/R,$$

which is assumed under the null hypothesis to have an F distribution with R and n - H (or n - H - R) degrees of freedom (see Shah, Holt, and Folsom, 1977).

This is a straightforward generalization of the conventional t-statistic in equation (3). Unfortunately, there is no analogous way to generalize our adjusted t_*-statistic in (6). Constructing a matrix analogue to equation (5) is not difficult. The stumbling block is determining the effective degrees of freedom for a putative "adjusted F-statistic."

The joint hypothesis $\mathbf{Q}\beta = \mathbf{h}$ can be viewed as the union R component hypotheses corresponding to the rows of $\mathbf{Q}$ and elements of $\mathbf{h}$. Let us denote the rth row of $\mathbf{Q}$ as $\mathbf{q}_r$ and the rth element of $\mathbf{h}$ as h_r. One can test each of the R component hypotheses in $\mathbf{Q}\beta = \mathbf{h}$ with its own adjusted t-test. Note that there is no guarantee that the respective t_*-statistics have the same effective degrees of freedom. This is why it is prohibitively difficult to determine the effective degrees of freedom for a putative adjusted F-statistic.

One practical way to test the joint hypothesis $\mathbf{Q}\beta = \mathbf{h}$ is through a Bonferroni procedure (Korn and Graubard, 1990, suggest this approach for a different reason). That is to say, one can test each of the R component null hypotheses using the appropriate adjusted t-test at the α/R level. If one component null hypothesis fails, say, $\mathbf{q}_r\beta$ is found significantly different from $\mathbf{h}_r$ at the α/R level for some r, then the joint null hypothesis, $\mathbf{Q}\beta = \mathbf{h}$, fails; if all R components "pass", then so does the joint null hypothesis.

The Bonferroni procedure is known to be conservative (that is, it does not reject null hypotheses as often as it should). An "improved Bonferroni" has been proposed by Simes (1986), who speculates that it is, if anything, still conservative "for a large family of multivariate distributions" when the component hypotheses are not independent. Unlike the Bonferroni, the Simes procedure is exact when the component hypotheses are

independent.

Let $p_{(1)}$, $p_{(2)}$, ..., $p_{(R)}$ be the ordered p-values for the R component hypotheses using the appropriate adjusted t-statistics. The Simes procedure rejects the joint null hypothesis, $\mathbf{Q\beta} = \mathbf{h}$, at the α level when $p_{(j)} \leq j\alpha/R$ for some j. The joint null hypothesis is accepted when $p_{(j)} > j\alpha/R$ for all j. It is easy to see that this procedure is less conservative than the original Bonferroni procedure which depends solely on whether $p_{(1)}$ is greater than α/R or not.

7 DISCUSSION

Let us return to the motivating example in section 2. It is not difficult to see that applying equation (5) to the linearization variance estimator, v_L, produces the exactly unbiased variance estimator, v_E. Assuming identically distributed errors within set A and calculating the effective degrees of freedom, F, with equation (11) yields 9.99. This is almost exactly one degree too many but clearly better than 97 or 99.

Repeated application of equation (12) on 10,000 simulated data sets constructed under the assumption that the ϵ_i in the motivating example are normal, independent, and identically distributed yielded an average f value of approximately 11.2 with a standard deviation of about 3.5. The average f value was greater than F due to the variability of the denominator of equation (12). By contrast, the average value of 1/f was roughly 0.100 ($\approx$ 1/9.99), as expected.

What this rather synthetic example shows is not so much how well the adjusted t-test works, but how misleading conventional design-based practice can be even with an apparently large sample size. The adjusted t-test, even when equation (12) is estimated from the sample, is clearly a giant step in the right direction.

As noted earlier, design-based techniques (sampling-weighted regression, the linearization variance estimator with stratification) also provide protection when the model in equation (1) fails, for example, when there are regressors missing from the equation. Unfortunately, this protection can not be addressed in the strictly model-dependent framework adopted here. It would be foolish, however, to expect the conventional design-based t-statistic to behave any better when the model in equation (1) fails than when it holds.

One potential problem of the test statistics suggested here when the model in (1) does not fail is that they may not be very powerful. Power can be lost by estimating regression coefficients with sampling weights, by not modelling the error structure, and by using a Simes-Bonferroni for testing joint hypotheses.

Returning to the motivating example can illustrate this point forcefully. If all the ϵ_i are assumed to be identically distributed, then the model-based v_M, which depends on the assumption of homoscedasticity, is unbiased and has 98 degrees of freedom, as compared to v_E with only 9. Often in practice, however, it

will be prudent to sacrifice power for robustness. One of the purposes of this paper was to provide a better measure of just how much power is lost using (modified) design-based methods when testing hypotheses about linear regression coefficients.

APPENDIX A: The Asymptotic Framework

Many of the results in this paper rely on the assumption that n, the number of primary sampling units in the sample, is large. (Formally, we should assume that there are infinite sequences of statistics -- $\{t_n\}$, $\{s_n^2\}$, etc. -- taking on values as n grows arbitrarily large.) If n is large, then so too must be M and m, the number of elements in the population and the sample, respectively. We will assume that $\max\{m_{hj}\}$ is bounded by a finite value, say $\bar{m}_0$. Thus, m is bounded by $\bar{m}_0 n$ and the number of nonzero elements in the block-diagonal matrix Σ_S is bounded by $\bar{m}_0^2 n$.

We have some flexibility concerning H. Either H can stay fixed as n grows arbitrarily large with the n_h/n ratios converging to fixed positive limits, or H/n can converge to a fixed positive limit with $\max\{n_h\}$ bounded.

Our concern in these appendices is with providing sufficient conditions for the results in the text to hold. Thus, the random variable x (formally, the infinite random sequence $\{x_n\}$) will be said to equal $O_P(n^{-\delta})$ when $|E(x^2)| < B/n^{2\delta}$ for some finite B. The random matrix $\mathbf{X}$ will be said to equal $\mathbf{O}_P(n^{-\delta})$ when each element x_{ij} in $\mathbf{X}$ satisfies $|E(x_{ij}^2)| < B/n^{2\delta}$. When x is not random, the P subscript on O is not needed. The same is true for $\mathbf{O}$.

The following two assumptions are reasonable given the structure that has been laid out:

(a) $\mathbf{C} = (\mathbf{X'WX})^{-1}\mathbf{X'W}$ exists and is $\mathbf{O}(1/n)$, and

(b) $E(\hat{\Sigma}_{hj}) = \Sigma_{hj} + \mathbf{O}(1/n)$
(recall that $\hat{\Sigma}_{hj} = \mathbf{r}_{hj}\mathbf{r}_{hj}'$).

Assumption (a) assures us that $\mathbf{Var}(\hat{\beta}_W) = \mathbf{C}\Sigma_S\mathbf{C}' = \mathbf{O}(1/n)$ (since there are m elements in the rows of $\mathbf{C}$ and no more than $\bar{m}_0^2 n$ non-zero elements in Σ_S).

The random variable $s^2 = \mathbf{qmse}_L\mathbf{q}$, where $\mathbf{q}$ is a row vector, can be rewritten as

$$s^2 = \sum_{h=1}^{H} (n_h/[n_h-1]) \sum_{j=1}^{n_h} (\mathbf{g}_{hj} - \mathbf{g}_h)\mathbf{r}_S\mathbf{r}_S'(\mathbf{g}_{hj} - \mathbf{g}_h)', \qquad (A1)$$

where $\mathbf{g}_{hj} = \mathbf{qCD}_{hj}$, $\mathbf{D}_{hj}$ is a diagonal matrix with 1's corresponding to the sampled elements of primary sampling unit hj and 0's elsewhere, and $\mathbf{g}_h = n_h^{-1}\Sigma\,\mathbf{g}_{hj}$ where the summation is across the j in h, as in the text.

Recall that $d = (s^2 - v^2)/v^2$. Now

$E(\mathbf{g}_{hj}\hat{\Sigma}_S\mathbf{g}_{hj}') = \mathbf{g}_{hj}\Sigma_S\mathbf{g}_{hj}' + O(n^{-3})$,
$E(\mathbf{g}_h\hat{\Sigma}_S\mathbf{g}_{hj}') = \mathbf{g}_h\Sigma_S\mathbf{g}_{hj}' + O(n^{-3})$,
and $E(\mathbf{g}_h\hat{\Sigma}_S\mathbf{g}_h') = \mathbf{g}_h\Sigma_S\mathbf{g}_h' + O(n^{-3})$.

Consequently, $E(s^2 - v^2) = O(n^{-2})$, and $E(d) = O(1/n)$. Moreover, since d can be put in the form of a linear combination of n independent $O(1/n)$ random variables, $d = O_P(n^{-1/2})$.

APPENDICES B AND C

Appendix B, which addresses the relative model biases of s^2 and s_*^2, and Appendix C, which establishes equations (8) through (12), are available from the author upon request.

DISCLAIMER

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

REFERENCES

Fuller, W.A. (1975). Regression analysis for sample survey. Sankhya C, **37**, 117-132.

Kish, L. & Frankel, M.R. (1974). Inference from complex samples. J. R. Statist. Soc. B, **36**, 1-37.

Korn, E.L. & Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: use of Bonferroni t Statistics. Amer. Statist. **44**, 270-276.

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

Rao, J.N.K. & Bellhouse, D.R. (1989). The history and development of the theoretical foundations of survey based estimation and statistical analysis. Amer. Statist. Assn. Proc. Sesq. Inv. Pap. Ses. 406-428.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. Biometrika **57**, 377-387.

Rust, K. (1987). Practical problems in sampling error estimation. Bul. Int. Stat. Inst. **52**, 3, 39-56.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. Biometrics **2**, 110-114.

Shah, B.V., Holt, M.M., & Folsom, R.E. (1977). Inference about regression models from sample survey data. Bul. Int. Statist. Inst. **47**, 43-57.

Simes, R.J. (1988). An improved Bonferroni procedure for multiple tests of significance. Biometrika **73**, 751-754.

Skinner, C.J. (1989). Domain means, regression, and multivariate analysis. In Analysis of Complex Surveys, eds. C.J. Skinner, D. Holt, & T.M.F. Smith, New York: John Wiley, pp. 59-88.

Wu, C.J.F., Holt, D., & Holmes, D.J. (1988). The effect of two Stage sampling on the F statistic. J. Amer. Statist. Assoc. **83**, 150-159.