Survey Methodology, December 1994
Vol. 20, No. 2, pp. 159–164
Statistics Canada

A Hypothesis Test of Linear Regression Coefficients with Survey Data

PHILLIP S. KOTT[1]

ABSTRACT

This paper discusses testing a single hypothesis about linear regression coefficients based on sample survey data. It suggests that when the design-based linearization variance estimator for a regression coefficient is used it should be adjusted to reduce its slight model bias and that a Satterthwaite-like estimation of its effective degrees of freedom be made. A very important special case of this analysis is its application to domain means.

KEY WORDS: Design-based; Domain mean; Effective degrees of freedom; Model-dependent; Probability order.

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-dependent 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-dependent theory) but the sample selection process. Rao and Bellhouse (1989) provides a useful summary of both design-based and model-dependent theory and of attempts to synthesize the two approaches.

The main concern here will be in the testing of a single hypothesis 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 *et al.* (1988), we will not explicitly model the error structure (except, perhaps, at a latter stage). Rather, we will focus our attention on a t-statistic calculated using the linearization variance estimator. That this variance estimator has desirable robustness properties from a model-dependent point of view has been demonstrated by Skinner (1989) and Kott (1991).

This paper will provide methods for reducing the model bias of the linearization variance estimator and for determining its effective degrees of freedom. A very important special case of this analysis is its application to the estimated variance of domain means and the difference of such means. Since the analysis in this paper is strictly model-dependent, the terms "bias" and "variance" will refer to model bias and model variance unless otherwise specified.

2. THE MODEL

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

$$y_M = X_M\beta + \epsilon_M, \tag{1}$$

where y_M is an $M \times 1$ vector of population values for the designated dependent variable;

X_M is an $M \times K$ matrix of population values for the K designated independent variables;

β is a $K \times 1$ vector of regression coefficients; and

ϵ_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 L strata. From each stratum h, n_h *distinct* clusters of elements are randomly sampled and denoted $u_{h1}, u_{h2}, \ldots, u_{hn_h}$. A random sample of m_{hj} elements is selected from each cluster hj. The clusters are also referred to as primary sampling units. There are $n = \sum n_h$ primary sampling units in the sample.

[1] Phillip S. Kott, National Agricultural Statistics Service, 3201 Old Lee Highway, Fairfax, VA 22030, U.S.A.

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. Observe that the sampling weights have been normalized so that if p_{hji} equals the sampling fraction, m/M, then w_{hji} would be unity.

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

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

where $\boldsymbol{y}_S$, for example, is the $m \times 1$ 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 $m_{hj} \times m_{hj}$ matrix $\boldsymbol{\Sigma}_{hj}$, which need not be diagonal. We assume that the $\boldsymbol{\epsilon}_{hj}$ are uncorrelated across primary sampling units, so that $\boldsymbol{\Sigma}_S$ is block diagonal.

The design-based estimator for $\boldsymbol{\beta}$ is the weighted least squares estimator:

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

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

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

The Taylor series or linearization estimator for the mean squared error of $\boldsymbol{b}_W$ (Shah *et al.* 1977) is

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

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

It should be noted that in their derivation of mse, Shah *et al.* 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 with*out* 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-dependent 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 $L > 1$, then there is an alternative to **mse** 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 ($L = 1$). Since the alternative can be expressed using equation (2), there is no need to treat it separately in the analysis that follows.

3. A CONVENTIONAL DESIGN-BASED *t*-STATISTIC

The estimator $\boldsymbol{b}_W$ is a K-vector. In this section we will be interested in the t-statistic used to test the univariate hypothesis that $\boldsymbol{q}\boldsymbol{\beta} = \Theta_0$ for some K element row vector $\boldsymbol{q} = (q_1, q_2, \ldots, q_K)$. The most common example of such an hypothesis addresses whether a particular element of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, say β_k, is zero. In this example, all of the q_t would be zero except q_k which would be 1; Θ_0 would also be zero.

If the model in (1) and the null hypothesis that $\boldsymbol{q}\boldsymbol{\beta} = \Theta_0$ are true, then

$$\Theta = (\boldsymbol{q}\boldsymbol{b}_W - \Theta_0)/\{\boldsymbol{q}\mathbf{Var}(\boldsymbol{b}_W)\boldsymbol{q}'\}^{1/2}$$

would be normally distributed with mean 0 and variance 1. If $\mathbf{Var}(\boldsymbol{b}_W)$ were known, the null hypothesis could be tested by comparing the statistic $\boldsymbol{\Theta}$ to a standard normal table. Unfortunately, $\mathbf{Var}(\boldsymbol{b}_W)$ must be estimated from the sample. Conventional design-based practice is to compare the statistic

$$t = (\boldsymbol{q}\boldsymbol{b}_W - \Theta_0)/(\boldsymbol{q}\boldsymbol{m}\boldsymbol{s}\boldsymbol{e}\boldsymbol{q}')^{1/2}, \qquad (3)$$

to a Student's t distribution with $n - L$ or $(n - L - K)$ degrees of freedom (see Shah *et al.* 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 $\boldsymbol{\Sigma}_S$ is block-diagonal. This will be done by investigating $s^2 = \boldsymbol{q}\boldsymbol{m}\boldsymbol{s}\boldsymbol{e}\boldsymbol{q}'$ as an estimator for $v^2 = \boldsymbol{q}\mathbf{Var}(\boldsymbol{b}_W)\boldsymbol{q}'$. First, s^2 will be adjusted to reduce its bias; then, a better determination of the adjusted estimator's effective degrees of freedom will be established.

4. THE MODEL BIAS OF s^2

The analysis to be conducted is asymptotic. Many of the results 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 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 $\bar{m}_0^2 n$.

The number of columns of X_S, K, is assumed to be fixed, but we have some flexibility concerning the number of strata, L. Either L can stay fixed as n grows arbitrarily large with the n_h/n ratios converging to fixed positive limits, or L/n can converge to a fixed positive limit with $\max\{n_h\}$ bounded.

Our concern here is with providing *sufficient* conditions for the subsequent analysis in the text to hold. The random variable ϕ (formally, the infinite random sequence $\{\phi_n\}$) will be said to be of probability order $n^{-\delta}$, *i.e.*, $\phi = O_P(n^{-\delta})$, when $|E(\phi^2)| < B/n^{2\delta}$ for some finite B. Similarly, the random matrix $\mathbf{\Phi}$ will be said to equal $\boldsymbol{O}_P(n^{-\delta})$ when each element ϕ_{ij} in $\mathbf{\Phi}$ satisfies $|E(\phi_{ij}^2)| < B/n^{2\delta}$. When ϕ is not random, the P subscript on O is not needed. The same is true for $\boldsymbol{O}$.

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

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

(2) $E(\hat{\Sigma}_{hj}) = \Sigma_{hj} + O(1/n)$, where $\hat{\Sigma}_{hj} = r_{hj}r_{hj}'$.

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

The variance of $\boldsymbol{qb}_W$ can be rewritten as $v^2 = \sum\sum v_{hj}/n^2$, where $v_{hj} = n^2 \boldsymbol{g}_{hj}\Sigma_S\boldsymbol{g}_{hj}$, $\boldsymbol{g}_{hj} = \boldsymbol{qCD}_{hj}$, and $\boldsymbol{D}_{hj}$ is a diagonal matrix with 1's corresponding to the sampled elements of primary sampling unit hj and 0's elsewhere. Similarly, $s^2 = \boldsymbol{qmseq}'$ can be rewritten as

$$s^2 = \sum_{h=1}^{L} (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (\boldsymbol{g}_{hj} - \boldsymbol{g}_h) r_S r_S' (\boldsymbol{g}_{hj} - \boldsymbol{g}_h)' \tag{4}$$

$$= \sum (n_h/[n_h - 1]) \sum [\boldsymbol{g}_{hj}\hat{\Sigma}_S\boldsymbol{g}_{hj}' - 2\boldsymbol{g}_h\hat{\Sigma}_S\boldsymbol{g}_{hj}' + \boldsymbol{g}_h\hat{\Sigma}_S\boldsymbol{g}_h'],$$

where $\boldsymbol{g}_h = \sum \boldsymbol{g}_{hj}/n_h$, the summation is across the j in h, and $\hat{\Sigma}_S = \sum\sum \boldsymbol{D}_{hj} r_S r_S' \boldsymbol{D}_{hj}$.

Both $\boldsymbol{g}_{hj}$ and $\boldsymbol{g}_h$ are $\boldsymbol{O}(1/n)$ because $\boldsymbol{C} = \boldsymbol{O}(1/n)$ and $\boldsymbol{D}_{hj}$ has a bounded number of non-zero values. Thus, $E(\boldsymbol{g}_{hj}\hat{\Sigma}_S\boldsymbol{g}_{hj}') = \boldsymbol{g}_{hj}\Sigma_S\boldsymbol{g}_{hj}' + O(n^{-3})$, $E(\boldsymbol{g}_h\hat{\Sigma}_S\boldsymbol{g}_{hj}') = \boldsymbol{g}_h\Sigma_S\boldsymbol{g}_{hj}' + O(n^{-3})$, and $E(\boldsymbol{g}_h\hat{\Sigma}_S\boldsymbol{g}_h') = \boldsymbol{g}_h\Sigma_S\boldsymbol{g}_h' + O(n^{-3})$. Consequently, $E(s^2 - v^2) = O(n^{-2})$.

Since $r_S = (I_m - \boldsymbol{XC})\epsilon_S$ and $E(\epsilon_S\epsilon_S') = \Sigma_S$, $E(r_S r_S') = \Sigma_S - \boldsymbol{XC}\Sigma_S - \Sigma_S\boldsymbol{C}'\boldsymbol{X}' + \boldsymbol{XC}\Sigma_S\boldsymbol{C}'\boldsymbol{X}'$. From equation (4), we can see that $E(s^2) = v^2 - R$, where $R = \sum (n_h/[n_h - 1]) \sum (\boldsymbol{g}_{hj} - \boldsymbol{g}_h)\boldsymbol{Z}(\boldsymbol{g}_{hj} - \boldsymbol{g}_h)'$ and $\boldsymbol{Z} = 2\boldsymbol{XC}\Sigma_S - \boldsymbol{XC}\Sigma_S\boldsymbol{C}'\boldsymbol{X}'$. Now $\boldsymbol{Z} = \boldsymbol{O}(1/n)$, because $\boldsymbol{C} = \boldsymbol{O}(1/n)$, $\boldsymbol{X}$ has a fixed number of columns, and the number of non-zero terms in any column of Σ_S is bounded. This implies $R = O(n^{-2})$. Thus, $-R/v^2$, the relative bias of s^2, is $O(1/n)$.

An alternative estimator for v^2 with a reduced relative bias is

$$s_*^2 = s^2/(1 - s^{-2}\hat{R}), \tag{5}$$

where

$$\hat{R} = \left\{ \sum_{h=1}^{L} (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (\boldsymbol{g}_{hj} - \boldsymbol{g}_h)\hat{\boldsymbol{Z}}(\boldsymbol{g}_{hj} - \boldsymbol{g}_h)' \right\},$$

and

$$\hat{\boldsymbol{Z}} = 2\boldsymbol{XC}\hat{\Sigma}_S - \boldsymbol{XC}\hat{\Sigma}_S\boldsymbol{C}'\boldsymbol{X}'.$$

In equation (5), $\hat{R}/s^2$ is used to estimate R/v^2. The variance estimator s_*^2 has been proposed here rather than the more obvious $s^2 + \hat{R}$ as *ad hoc* compensation for the slight relative bias of $\hat{R}$ as an estimator of R.

5. THE RELATIVE VARIANCE OF THE VARIANCE ESTIMATOR

Let $e_{hj} = n\boldsymbol{g}_{hj}\epsilon_S$ so that $\mathrm{Var}(e_{hj}) = v_{hj}^2$, and recall that $v^2 = \sum\sum v_{hj}^2/n^2$. If $\hat{e}_{hj} = n\boldsymbol{g}_{hj}r_S$, then the random variable s^2 can be re-written as

$$s^2 = n^{-2} \sum_{h=1}^{L} (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (\hat{e}_{hj} - \hat{e}_h)^2$$

$$= n^{-2} \sum (n_h/[n_h - 1]) \{ \sum (e_{hj} - e_h)^2 - (\boldsymbol{g}_{hj} - \boldsymbol{g}_h)\boldsymbol{A}(\boldsymbol{g}_{hj} - \boldsymbol{g}_h)' \},$$

where $\boldsymbol{A} = 2\boldsymbol{XC}e_S e_S' - \boldsymbol{XC}e_S e_S'\boldsymbol{C}'\boldsymbol{X}'$. It is now possible to show that

$$s_*^2 = n^{-2} \sum_{h=1}^{L} (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (e_{hj} - e_h)^2 + O_P(n^{-5/2}).$$

Consider a random variable with a χ^2 distribution with F degrees of freedom. Its relative variance is $2/F$. This suggests a Satterthwaite-like determination of the effective degrees of freedom of s_*^2 (see Satterthwaite 1946); namely,

$$F = \frac{(nv)^4}{\sum_{i=1}^{L}\left\{\sum_{j=1}^{n_h} v_{hj}^4 + \sum_{j'\neq j} v_{hj}^2 v_{hj'}^2/(n_h-1)^2\right\}}, \quad (6)$$

which is approximately 2 divided by the relative variance of s_*^2 (since $s_*^2 \approx n^{-2}\sum_h\{\sum_j e_{hj}^2 + \sum_{j'\neq j} e_{hj}e_{hj'}/(n_h-1)\}$).

What is being recommended here is that one tests whether $\boldsymbol{q\beta} = \Theta_0$ by assuming under the null hypothesis that

$$t_* = (\boldsymbol{q}\boldsymbol{b}_W - \Theta_0)/s_*, \quad (7)$$

has a Student's t distribution with F degrees of freedom, where F is determined using equation (6) and making some assumptions about the v_{hj}. Let us call this test the *adjusted t-test*.

6. A SIMPLE EXAMPLE

Consider a simple random sample of n units, n_1 of which are in a subset of the sample denoted by A and n_2 in the complement denoted $\bar{A}$. Let y_i be the observed value for unit i. Suppose the following linear model holds:

$$y_i = d_i\beta_1 + (1-d_i)\beta_2 + \epsilon_i, \quad (8)$$

where $d_i = 1$ is unit i is in set A, and 0 if i is in $\bar{A}$; and the ϵ_i are independent normally distributed random variables.

Assuming homoscedastic errors, both the model-dependent and design-based regression estimator for β_1 is the simple domain mean, $\bar{y}_A = \sum_{i\in A} y_i/n_1$. The linearization estimator for the variance of this estimator is simply $v_L = (n/[n-1])\sum_{i\in A}(y_i - \bar{y}_A)^2/n_1^2$. (It should be noted that when a domain mean is viewed as an analytic parameter, its variance requires no finite population correction; see Fuller 1975).

This linearization estimator, v_L, differs from model-dependent variance estimator: $v_M = [\sum_{i\in A}(y_i - \bar{y}_A)^2 + \sum_{i\in\bar{A}}(y_i - \bar{y}_{\bar{A}})^2]/[n_1(n-2)]$. The advantage of v_L is that, unlike v_M, it is asymptotically unbiased under the model even when the ϵ_i are heteroscedastic. This 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 $n_1 = 10$, the relative bias of v_L is approximately 10%. We can see this by noting that $v_E = \sum_{i\in A}(y_i - \bar{y}_A)^2/(n_1[n_1-1]) = ([n-1]/n)(n_1/[n_1-1])v_L$ is exactly unbiased.

Continuing the example: 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.

Applying equation (5) to the linearization variance estimator, v_L, produces a variance estimator virtually identical to v_E (since $\hat{R} = [v_L/n_1][1 - n_1/n]$, s_*^2 differs from v_E by only 0.1%). Assuming identically distributed errors within sets A and $\bar{A}$, calculating the effective degrees of freedom, F, with equation (6) yields 9.99. This is almost exactly one degree too many but clearly better than 97 or 99.

A natural hypthesis to test is whether the domain means, β_1 and β_2, in equation (8) are equal. In other words is $\beta_1 - \beta_2 = \Theta_0 = 0$? Assuming that all units have the same variance, the adjusted t statistic is

$$t^* = \frac{\sum_{i\in A} y_i/n_1 - \sum_{i\in\bar{A}} y_i/n_2}{(1 - s^{-2}\hat{R})^{1/2}s},$$

where

$$s^2 = [n/(n-1)][\sum_{i\in A}(y_i - \bar{y}_A)^2/n_1^2 + \sum_{i\in\bar{A}}(y_i - \bar{y}_{\bar{A}})^2/n_2^2],$$

and

$$\hat{R} = [n/(n-1)][(\sum_{i\in A}(y_i - \bar{y}_A)^2/n_1^3)(1 - n_1/n) + (\sum_{i\in\bar{A}}(y_i - \bar{y}_{\bar{A}})^2/n_2^3)(1 - n_2/n)].$$

To calculate the effective degrees of freedom for $ns^2/(n-1)$ - and thus t_* - using equation (6), note that $L = 1$, and $v_i \propto 1/n_1^2$ for $i \in A$ while $v_i \propto 1/n_2^2$ for $i \in \bar{A}$. As a result,

$$F = \frac{(1/n_1 + 1/n_2)^2}{(1/n_1^3 + 1/n_2^3 + [\{1/n_1 + 1/n_2\}^2 - 1/n_1 - 1/n_2])/n^2},$$

which is 12.3 when $n_1 = 10$ and $n_2 = 90$. The actual degrees of $ns^2/(n-1)$ (*i.e.*, 2 divided by its relative variance) is reasonably close, 11.1 (the relative variance of $ns^2/(n-1)$ is $2[(n_1-1)/n_1^4 + (n_2-1)/n_2^4]$ divided by $[(n_1-1)/n_1^2 + (n_2-1)/n_2^2]^2$).

What this synthetic example principally shows is how misleading conventional design-based practice can be even with an apparently large sample size. The adjusted t-test is clearly a giant step in the right direction.

It is tempting to try to avoid making an assumption about the v_{hj} and to estimate F with

$$f = \frac{(ns_0)^4 - \sum_{}^{L} \sum_{}^{n_h} 2s_{hj}^4/3}{\sum_{i=1}^{L} \left\{ \sum_{j=1}^{n_h} s_{hj}^4/3 + \sum_{j' \neq j} s_{hj}^2 s_{hj'}^2/(n_h - 1)^2 \right\}}, \quad (9)$$

where $s_{hj}^2 = n^2(g_{hj} r_S)^2$. Although f is a consistent estimator of F, its use can produce misleading results as we shall see.

Repeated application of equation (9) on 10,000 simulated data sets constructed under the assumption that the ϵ_i in equation (8) are normal, independent, and identically distributed yielded an average f value for the variance of $\bar{y}_A$ of approximately 11.2 with a standard deviation of about 3.5. In addition to its variability, the average f value is greater than F. This is due to the denominator of equation (9) itself being a random variable. It happens that the value of $1/f$ is roughly 0.100 ($\approx 1/9.99$), as expected. Thus, even though the use of f in equation (9) may seem appealing, it is not recommended.

7. ANOTHER EXAMPLE OF A DOMAIN MEAN

Faced with the simple example of the last section, most design-based statisticians would simply treat the units sampled from set A as an independent simple random sample. The linearization and model-dependent variance estimator would then coincide. In practice, however, samples often involve clustering, stratification, and unequal probabilities of selection. When the domain of interest is not a design stratum, it usually becomes impossible to separate out the domain's sampled elements (which need not be primary sampling units) and treat them as an independent random sample.

An example of such a complex sample is the 1985 Continuing Survey of Intakes by Individuals (CSFII). This was a stratified, multistage survey of the dietary intakes of women from 19 to 50 years of age and children from 1 to 5. There were roughly 140 women in the sample who described themselves as black and 1,150 who described themselves as white.

Assuming that a dietary intake value for each individual was independent and identically distributed, values of the relative variance of the linearization variance estimator (R/s^2 from equation (5)) and its effective degrees of freedom (F from equation (6)) were calculated for the two race domains. The relative bias for white women was .003, while the effective degrees of freedom were 48.1. For black women, the relative bias was 0.026, and the effective degrees of freedom 10.1. Thus, even with a fairly large sample size, the effective number of degrees of freedom for black women was relatively small. The conventional determination of degrees of freedom was around 60 (120 PSU's minus 60 design strata).

8. DISCUSSION

As pointed out earlier, the use of design-based techniques can often provide protection when the model in equation (1) fails. Unfortunately, this protection can not be addressed in the strictly model-dependent framework adopted here. It would be unrealistic, however, to expect a 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 modified design-based test statistic suggested here occurs when the model in equation (1) does *not* fail: it may not be very powerful. Power can be lost by estimating regression coefficients with sampling weights and by not modelling the error structure directly.

This loss of power is due to the original design-based formulation and not to our modification of it. In fact, s_*^2 is a design consistent estimator of the design mean squared error of b_W whenever s^2 is. This is because $\hat{R}/s^2$ in equation (5) is also $O_P(1/n)$ from a design-based point of view assuming that the first stage of sampling is conducted *with* replacement.

Returning to the simple example of Section 6 can illustrate the issue of power forcefully. The model-dependent and design-based estimates are the same. If all the ϵ_i are assumed to be identically distributed, then the model-dependent variance estimator, v_M, which depends on the assumption of homoscedasticity, is unbiased and has 98 degrees of freedom. The adjusted design based variance estimator is also virtually unbiased, but it has only 9 degrees of freedom.

Often in practice, it will be prudent to sacrifice power for robustness. When that is the case, equation (6) provides an attractive method of measuring how much power may be lost using a modified design-based t-test (equation (7)) when the assumptions of the model are, in fact, correct. Furthermore, the equation lends itself to sensitivity analyses in which the effects of alternative assumptions about the v_{hj} can be evaluated.

ACKNOWLEDGEMENTS

The author would like to thank the staff of the Beltsville Human Nutrition Research Center for its support of this research and an associate editor and his (her) referees for their helpful comments.

REFERENCES

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

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

RAO, J.N.K., and BELLHOUSE, D.R. (1989). The history and development of the theoretical foundations of survey based estimation and statistical analysis. *American Statistical Association Proceedings Sesquincentennial Invited Paper Sessions*, 406-428.

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

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

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.

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

WU, C.J.F., HOLT, D., and HOLMES, D.J. (1988). The effect of two stage sampling on the F statistic. *Journal of the American Statistical Association*, 83, 150-159.