

LINEAR REGRESSION IN THE FACE OF SPECIFICATION ERROR: A MODEL-BASED EXPLORATION OF RANDOMIZATION-BASED TECHNIQUES

P. S. Kott¹

ABSTRACT

Much of statistics is predicated on the paucity of data. This drives a constant search for statistically efficient techniques and parsimonious models. Survey sampling is one branch of statistics where sample sizes are almost always large. As a result, efficiency considerations are usually secondary to questions of robustness. That is especially true when estimating the parameters of a linear model, which may include the use of instrumental variables. Randomization-based techniques developed for analysing survey data can be shown to provide protection against the possibility that the independent variables of the linear model are misspecified. Moreover, statistical inferences about model parameters need not rely on often dubious assumptions about the error structure of the model. Few survey statisticians recognize the costs of using randomization-based techniques, however. These costs are measured for parameter estimates using real survey data. The results are somewhat surprising.

KEY WORDS: Effective degrees of freedom; Extended linear model; Instrumental variable; Primary sampling unit; Putative missing regressor; Stratum.

RÉSUMÉ

La majorité des statistiques sont contraintes par la pénurie de données. Ceci nous amène à chercher constamment des techniques statistiquement efficaces et des modèles qui nécessitent peu de données. L'échantillonnage est un domaine des statistiques où la taille des échantillons est presque tout le temps grande. En conséquence, les considérations d'efficacité sont souvent secondaires aux questions de robustesse. Ceci est particulièrement vrai lors de l'estimation des paramètres dans les modèles linéaires des variables contributives. Il peut être démontré que les techniques utilisant la répartition au hasard ou aléatoire, qui ont été développées pour analyser les données d'enquête apportent une protection contre la possibilité que les variables indépendantes soient mal spécifiées dans le modèle linéaire. De plus, l'inférence statistique des paramètres des modèles ne doivent pas nécessairement se baser sur des hypothèses douteuses au sujet de la structure de l'erreur dans le modèle. Malheureusement, peu de statisticiens d'enquête réalisent le plein coût de l'utilisation de ces techniques. Ces coûts associés à l'estimation du paramètre sont mieux mesuré en utilisant les données réelles. Les résultats sont assez surprenants.

MOTS CLÉS: Degrés de liberté effectifs; modèle linéaire étendu; variable contributive; unité primaire d'échantillonnage; variable explicative putative; strate.

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 *randomization-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 randomization-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. Fuller

¹ Phillip S. Kott, USDA/NASS, Fairfax, VA 22030, U.S.A.

(1975, 1984) has proposed replacing the focus of randomization-based linear regression from a fixed population to an unspecified - and not necessarily linear - model generating the population values. This approach does not always extend well (see, for example, Kott 1991b) and will be treated here only in passing.

Shah *et al.* (1977) discuss standard randomization-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. Kott (1991a) proposes an expanded framework for the linear model under which Shah *et al.*'s randomization-based techniques have useful model-based properties when there are missing regressors and/or the error variance matrix is only vaguely specified.

The price one pays for using randomization-based techniques is loss of efficiency, not only in the estimated regression coefficient itself but in the estimation of its variance. In a later paper, Kott (1994) provides a strictly model-based means of measuring this second source of loss.

Section 2 of this paper reviews estimation under the conventional linear model. Section 3 discusses the properties of randomization-based estimation techniques under the extended linear model in Kott (1991a). It also addresses a method for removing the bias of the so-called linearization variance estimator under a particular model and of calculating an estimated regression coefficient's effective degrees of freedom. Section 4 discusses the use of instrumental variables in linear regression under the extended model. Section 5 applies the effective degrees of freedom calculations to data from a real survey with some surprising results. Section 6 provides a discussion.

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:

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

where $\mathbf{y} = (y_1, \dots, y_M)'$ is an $M \times 1$ vector of population values for a dependent variable; \mathbf{X} is an $M \times K$ matrix of population values for K independent variables or regressors; $X_{(k)}$ denotes one of the regressor variables in \mathbf{X} (*i.e.*, a column of \mathbf{X}); $\boldsymbol{\beta}$ is a $K \times 1$ vector of

regression coefficients; and $\boldsymbol{\epsilon}$ is an $M \times 1$ vector of disturbances or errors independent of \mathbf{X} and satisfying $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and $\text{Var}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 \mathbf{I}_M$.

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

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

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 $\boldsymbol{\epsilon}$.

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

$$b_{OLS} = (\mathbf{X}'\mathbf{S}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{S}\mathbf{y}),$$

where \mathbf{S} is an $M \times M$ diagonal matrix of 0's and 1's. The i -th diagonal of \mathbf{S} is 1 if and only if the i -th element of the population is in the sample.

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

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

3. RANDOMIZATION-BASED TECHNIQUES AND THE EXTENDED LINEAR MODEL

Let \mathbf{P} be a $M \times M$ diagonal matrix whose i 'th diagonal is the probability element i was selected for the sample. We can call $\mathbf{W} = (m/M)\mathbf{S}\mathbf{P}^{-1}$ 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 under orthodox randomization-based theory, the weighted regression estimator,

$$b_w = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{y}), \quad (3)$$

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

3.1 The Extended Linear Model

Kott (1991a) observes that 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:

$$y = X\beta + z + \epsilon, \quad (4)$$

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

Under mild conditions on the survey variables and sampling design, b_w is nearly - that is, asymptotically - unbiased under the model for large m because $\text{plim}_{m \rightarrow \infty} X'Wz/m = O_p(1/\sqrt{m})$. The same cannot be said for b_{OLS} unless $\lim_{M \rightarrow \infty} X'Pz/m = 0$, which in practical terms means that the probabilities of selection are unrelated to the missing regressors.

3.2 Variance Estimation

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

Let n_h be the number of PSU's selected from stratum h and $n = \sum n_h$. We can rewrite b_w in equation (3) as

$$b_w = \sum_{h=1}^H \sum_{j=1}^{n_h} (X'WX)^{-1} X'WD_{hj}y,$$

where D_{hj} is a $M \times M$ diagonal matrix of 1's and 0's such that the i 'th diagonal of D_{hj} is 1 if and only if the i 'th member of y corresponds to an element in PSU hj .

Let $g_{hj} = (X'WX)^{-1} X'WD_{hj}(y - Xb_w)$ for each PSU hj . The so-called *linearization variance estimator* for b_w is the matrix:

$$V = \sum_{h=1}^H \frac{n_h}{n_{h-1}} \left[\sum_{j=1}^{n_h} g_{hj} g_{hj}' - \frac{1}{n_h} \left(\sum_{j=1}^{n_h} g_{hj} \right) \left(\sum_{j=1}^{n_h} g_{hj} \right)' \right]. \quad (5)$$

This is the estimator computed by the SUDAAN (Shah *et al.*, 1991) software package when the design is specified as being *with-replacement* in the first stage and independent across PSU's thereafter. PC CARP

(Fuller *et al.*, 1986) scales V by $\{(m-1)/(m-K)\}$, which is asymptotically unity.

Returning to the extended model in equation (4) and assuming that $z = \mathbf{0}$ (so that b_w is exactly model unbiased), V is a nearly unbiased estimator for the variance of b_w under mild conditions so long as the following holds:

Condition (v): $E(\epsilon_i \epsilon_a) = 0$ when i and a are elements from different PSU's and bounded otherwise.

A more efficient variance estimator under the same model is

$$V' = \frac{n}{n-1} \sum_{h=1}^H \sum_{j=1}^{n_h} g_{hj} g_{hj}', \quad (6)$$

which equals V when $H=1$ and $\sum \sum g_{hj} = \mathbf{0}$. Both V and V' rely on the fact that

$$\begin{aligned} m(X'WX)^{-1} X'WD_{hj}(y - X\beta) \\ &= m(X'WX)^{-1} X'WD_{hj}\epsilon \\ &\approx mg_{hj} \end{aligned}$$

are (nearly) independent random variables under the model with mean $\mathbf{0}_K$.

3.3 The Accuracy of the Linearization Variance Estimator

When z is not identically zero, b_w is not model unbiased. Rather it has a model bias,

$$a = (X'WX)^{-1} X'Wz,$$

of probability order $O_p(1/\sqrt{m})$. This model variance of b_w is

$$\Lambda = (X'WX)^{-1} X'WE(\epsilon \epsilon') W'X(X'WX)^{-1},$$

which is $O(1/n)$ under mild conditions. Thus, if m/n converges asymptotically to a positive constant, the model bias of b_w is *not* an asymptotically ignorable component of its model mean squared error. Consequently, we need to make assumptions about the size of the putative missing regressor vector, z . In particular, we assume here that z is such that $aa' \ll \Lambda$, and $D_{hj}zz'D_{hj} \ll D_{hj}E(\epsilon \epsilon')D_{hj}$ for all hj . As a result, z has no appreciable impact on the model mean squared error/variance of b_w or on its estimation. It may still, however, have a consequence on the model mean squared error of b_{OLS} , because the model bias of that estimator can be $O_p(1)$. Thus, the impact on b_{OLS} of

a very small z need not tend to zero as the sample size grows asymptotically large.

Even when z is assumed to be identically zero, V in equation (5) has a small model bias of asymptotic order $O(1/n)$. This is because

$$g_{hj} = (X'WX)^{-1}X'WD_{hj}(y - Xb_w)$$

does not exactly equal $(X'WX)^{-1}X'WD_{hj}\epsilon$; rather, $g_{hj} = (X'WX)^{-1}X'WD_{hj}[I_M - X(X'WX)^{-1}X'W]\epsilon$. As a result, the model bias of V is

$$E(V - \Lambda) = -\sum_h [(\sum_j (g_{hj} Z g_{hj}') - n_h g_h Z g_h')],$$

where $g_h = \sum_j g_{hj}/n_h$, and

$$Z = 2X(X'WX)^{-1}X'WE(\epsilon\epsilon') - X(X'WX)^{-1}X'WE(\epsilon\epsilon')W'X(X'WX)^{-1}X'$$

Let us redirect our focus from the entire estimated regression coefficient vector, b_w , to a particular member of b_w . In particular, consider $b_o = qb_w$, where q is a row vector with one member equal to unity and the rest equal to zero. The linearization variance estimator for b_o is $v_o = qVq'$.

Suppose $T = E(\epsilon\epsilon')$ were known up to a constant, then an unbiased estimator for the model variance of b_o is

$$v_u = qVq'R, \quad (7)$$

where $R = (a - c)/a$,

$$a = q(X'WX)^{-1}X'WTW'X(X'WX)^{-1}q',$$

$$c = \sum_h [(\sum_j (qg_{hj} Qg_{hj}' q') - n_h qg_h Qg_h' q')], \text{ and}$$

$$Q = 2X(X'WX)^{-1}X'WT - X(X'WX)^{-1}X'WTW'X(X'WX)^{-1}X'.$$

Since R is asymptotically unity, v_u is nearly unbiased no matter what $E(\epsilon\epsilon')$ is as long as condition (v) holds; moreover, v_u is exactly unbiased when $E(\epsilon\epsilon')$ is a multiple of T . The use of the equation (7) to adjust for the bias in the linearization variance estimation differs from the suggestion in Kott (1994). Instead, it is a generalization of an approach proposed earlier in Kott (1989).

Let $\xi_{hj} = nqg_{hj}\epsilon$ and $\text{Var}(\xi_{hj}) = v_{hj}$. The model variance of b_o with this notation is $\sum \sum v_{hj}/n^2$. If $e_{hj} = nqg_{hj}r_s$, then the linearization variance estimator can be re-written as

$$v_o = n^{-2} \sum_{h=1}^L (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (e_{hj} - e_h)^2$$

$$= n^{-2} \sum (n_h/[n_h - 1]) \sum (\xi_{hj} - \xi_h)^2 [1 + O_\epsilon(1/\sqrt{n})]$$

$$= v_u [1 + O_\epsilon(1/\sqrt{n})].$$

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 $v_o \approx v_u$; namely,

$$F = \frac{(\sum_{h=1}^L \sum_{j=1}^{n_h} v_{hj})^2}{\sum_{h=1}^L \{ \sum_{j=1}^{n_h} v_{hj}^2 + \sum_{g \neq j} v_{hj} v_{gh} / (n_h - 1)^2 \}}, \quad (8)$$

which is approximately 2 divided by the relative variance of $v_o \approx n^{-2} \sum_h \{ \sum_j \xi_{hj}^2 + \sum_{g \neq j} \xi_{hj} \xi_{hg} / (n_h - 1) \}$, assuming $E(\xi_{hj}^4) = 3v_{hj}^2$ - the fourth central moment of the normal distribution.

In order to use equation (8) in practice, we need to provide relative values for the v_{hj} . Sufficient for this is the assumption that $E(\epsilon\epsilon')$ is known up to a constant. Although it is tempting to try to estimate F by replacing the v_{hj} with e_{hj}^2 , Kott (1994) provides a simple example illuminating the problem with doing so (the result is very unstable).

4. INSTRUMENTAL VARIABLES

There are other estimators for β in equation (4) based on the entire population rather than B in equation (2). In particular, if G is an $M \times K'$ matrix such that $K' \geq K$, $E(\epsilon G_{(k)}') = \mathbf{0}_{M \times M}$, where $G_{(k)}$ denotes a column of G , and $\lim_{M \rightarrow \infty} G'z/M = O_{k'}$, then

$$B_G = [X'G(G'G)^{-1}G'X]^{-1}X'G(G'G)^{-1}G'y$$

is a nearly model unbiased estimator for β . The columns of G are called *instrumental variables*. We will focus in this section on the special case where $K' = K$, and B_G collapses to

$$B_G = (G'X)^{-1}G'y. \quad (9)$$

Armed with only survey data, we cannot determine B_G . Nevertheless, following the reasoning of the last section, one can see that the randomization-based estimator for B_G , namely,

$$b_G = (G'WX)^{-1}G'Wy \quad (10)$$

is nearly model unbiased under mild conditions provided the assumption that $\lim_{M \rightarrow \infty} G'z/M = \mathbf{0}_K$ holds. The reason we focus on this particular assumption will be made clear shortly.

When is instrumental regression used in practice? The most common application is when $K=1$ and X does not equal the M -vector of 1's, $\mathbf{1}_M$. If G in equation (10) is set to $\mathbf{1}_M$, the resultant scalar estimator, $b_{G=1}$, is generally preferred over b_w . Observe that $b_{G=1} = \sum w_i y_i / \sum w_i x_i$ is the weighted ratio estimator, while $b_w = \sum w_i x_i y_i / \sum w_i x_i^2$.

In many practical situations, b_w and $b_{G=1}$ – or more to the point, B_w and $B_{G=1}$ – are estimators of the same thing. This is not always the case, however. For example, suppose that the mechanism generating the population values produces values satisfying $y_i = \alpha x_i^\alpha$, where $\alpha \neq 1$, and the x_i are uniformly distributed on $(0,1)$. It is easy to show that B_w and $B_{G=1}$ do not converge as the population grows arbitrarily large.

The problem is that it is possible to describe the population with equation (4) such that $\sum z_i/M \rightarrow 0$ or such $\sum x_i z_i/M \rightarrow 0$, but not both. One reason for preferring $b_{G=1}$ over b_w is that the assumption that the putative missing regressor capturing the effects of all the missing regressors averages zero asymptotically is more appealing than the assumption that its x -weighted mean is asymptotically zero.

In a multivariate setting where X contains a column of 1's or the equivalent (*i.e.*, $X\lambda = \mathbf{I}_M$ for some λ), the assumption that z is such that $\lim_{M \rightarrow \infty} X'z/M = \mathbf{0}_K$ is not unreasonable. In this context, an additional – or alternative – assumption that $\lim_{M \rightarrow \infty} G'z/M = \mathbf{0}_K$ should not be made lightly.

Another reason for using instrumental variables in regression is because the true X matrix – or at least one regressor within X – is measured with a random, but unbiased, error. Let X^* denote the matrix of true values for the regressors and X the matrix of measured variables. It is common to assume that the rows of $U = X - X^*$ are uncorrelated, although for our purposes it suffices that condition (v) holds for the members of $\epsilon = \epsilon^* + (X^* - X)\beta$. Let the matrix G contain those columns of X that are measured without error but replace the rest in such a way that assures $\text{plim}_{M \rightarrow \infty} G'U/M = \mathbf{0}_{K \times K}$.

It is not hard to see that given such a G , B_G in equation (9) is a nearly unbiased estimator for β under the model:

$$y = X^*\beta + z + \epsilon^*, \quad (4')$$

as long as $E(\epsilon^*G_{(k)}') = \mathbf{0}_{M \times M}$ for all k , and $\text{plim}_{M \rightarrow \infty} G'z/M = \mathbf{0}_K$; “plim” rather than “lim” to allow the possibility that one of the added columns to G is random. Under mild conditions on the survey variables and sampling design b_G in equation (10) is likewise nearly model unbiased.

When X contains a variable measured with error, one common practice is to replace that column in G with a second variable measured with an independent error. For example, suppose y is a vector of values measuring the cholesterol levels for a population of individuals, and a column of X^* is a vector of the corresponding individuals' usual daily intakes of dietary fat. This latter vector is unavailable for the sample. In its place, we have two records of each sampled individual's one-day fat intake taken on two separate days.

For simplicity, assume that we can treat each intake record as an unbiased and independent (but not error-free) estimate of the corresponding individual's usual daily intake of fat. By putting one vector of one-day intakes in X and the other in G , b_G can become a nearly model unbiased estimator for β , in equation (4'). Observe, however, that since G must be linearly unrelated in the limit to the missing regressor vector, z (*i.e.*, $\text{plim}_{M \rightarrow \infty} G'z/M = \mathbf{0}_K$), and the columns of G need to be uncorrelated with ϵ^* (*i.e.*, $E(\epsilon^*G_{(k)}') = \mathbf{0}_{M \times M}$), the one-day intakes in G should be not measured on the same day as the corresponding individuals' cholesterol level. The use of instrumental variable regression to handle problems where usual daily dietary intakes are among the explanatory variables was suggested to me by Wayne Fuller. His EV Carp software (Schnell *et al.*, 1988) handles instrumental variable regression with survey data. The randomization-based theory supporting its use, in general, has no documentation as far as I know. For the particular example described above, arguments in Fuller (1975) – where the limit of B in equation (2) as the population grows arbitrarily large is essentially the target of estimation – apply in a fairly straightforward manner. This is because $\text{plim}_{M \rightarrow \infty} G'X/M = \lim_{M \rightarrow \infty} X^*X^*/M$, and $\text{plim}_{M \rightarrow \infty} G'y/M = \lim_{M \rightarrow \infty} X^*y/M$ in this context. These equalities do not hold for more general G , however. To estimate the model variance of b_G , conditioned on G when the latter is random, it is straightforward to show that equations (5), (6), and (8) still apply with g_{hj} redefined as

$$g_{hj} = (G'WX)^{-1}G'WD_{hj}(y - XB_G).$$

For equation (7) to apply, a becomes

$$a = q(G'WX)^{-1}G'WTW'G(X'WG)^{-1}q'$$

and Q becomes

$$Q = 2X(G'WX)^{-1}G'WT - X(G'WX)^{-1}G'WTW'G(X'WG)^{-1}X'$$

It should be noted that even when T specified $E(\epsilon\epsilon')$ correctly up to a constant, v_U in equation (7) will not be exactly model unbiased when X contains measurement error. This is because $b_G = (G'WX)^{-1}G'Wy$ must be a linear combination of random variables for the unbiasedness of v_U to hold, and it is not when X is random.

5. AN EXAMPLE

In this section, we apply equation (7) and (8) to real survey data with some surprising results. The 1994 Continuing Survey of Food Intake by Individuals (CSFII) is a stratified, multi-stage survey of dietary intake conducted by Westat for the US Department of Agriculture (USDA, Agricultural Research Service, 1996). For variance estimation purposes, there are 43 strata with two sampled PSU's in each. Sample weighting reflects the original probabilities of selection (certain population subdomains such as small children and individuals from low income households were sampled at relatively high rates) and adjustments for nonresponse and under coverage.

We restrict our attention to the 1046 women in the 1994 CSFII providing at least one day of dietary intake information who described themselves as being either white or black and between the ages 30 and 69. The 149 black women in this sample came from 49 different PSU's. 19 strata had sampled black women in both their PSU's. The 118 women aged 65 to 69 came from 58 PSU's. 20 strata had sampled women from this age group in both their PSU's.

The dependent variable analyzed was each individual's total fat intake for the first reported day of intake. Three linear models were explored. Model 1 was a simple regression estimator with two exhaustive and mutually exclusive dummies. Model 2 was the identical model reparameterized as an intercept and a single dummy. Model 3 added six covariates for the intake day-of-the week and 11 for the intake month to Model 2. Table 1 displays results when the single dummy in Model 2 was black and the second dummy in Model 1 was white. Estimated linearization standard errors for the black regression coefficient under each

model were calculated by taking the square root of v_o in equation (5). Bias-corrected standard errors (using equation (7)) and effective degrees of freedom (using (8)) were also calculated assuming that fat intakes across sampled individuals were uncorrelated and homo-scedastic (*i.e.*, $T = \sigma^2 I_M$) – the assumptions supporting OLS. The results for the three models are labelled the “standard approach” in the table.

Table 1. Standard Error and Effective Degrees of Freedom Estimates for the Dummy Regression Coefficient: BLACK

Model	Estimates		
	Linearization Standard Error $\sqrt{v_o}$	Corrected Standard Error $\sqrt{v_u}$	Effective Degrees of Freedom
		(with $T = \sigma^2 I_M$)	(with $T = \sigma^2 I_M$)
Standard Approach			
1	2.98	3.15	2.43
2	2.91	3.03	2.84
3	2.61	2.71	4.40
Ignoring the Stratification			
1	3.52	3.91	2.73
2	3.57	3.96	3.21
3	3.25	3.67	5.46
Ignoring the Stratification and the Clustering			
1	3.50	3.61	6.54
2	3.68	3.79	7.76
3	3.42	3.60	14.60

Model 1: FAT = β_1 BLACK + β_2 WHITE + ERROR
 Model 2: FAT = $\beta_0 + \beta_1$ BLACK + ERROR
 Model 3: FAT = $\beta_0 + \beta_1$ BLACK +
 DUMMIES FOR DAY-OF-THE-WEEK
 and MONTH OF INTAKE + ERROR

It is surprising how few effective degrees of freedom there are for the estimated standard error of the black regression coefficient under each of the models. In an attempt to increase the degrees of freedom, stratification was ignored in variance estimation (*i.e.*, equation (6) was used). The increase in estimated effective degrees of freedom is quite modest, much less than *ad hoc* speculation would suggest.

Going further and ignoring the clustering and treating each individual as a PSU increases the effective degrees of freedom less than three-fold. Even though there are 148 black women in the sample,

the estimated effective degrees of freedom for the standard error of the black regression coefficient is never as large as 15, even with 17 covariates! The problem appears to be the CSFII survey weights. Two of the black women lived in an apartment complex that was unexpected by survey planners. To save costs, residents of this complex were sampled at a much lower rate than elsewhere in the country. As a result, these two women have weights over nine times the size of any other woman in the sample.

Table 2 parallels Table 1 exactly except that the sample weights are ignored (*i.e.*, the matrix W is replaced by I_M in equations (5) through (8)). The estimated effective degrees of freedom increase dramatically. Moreover, the gains from ignoring the stratification and ignoring the clustering are much more pronounced.

Table 2. Standard Error and Effective Degrees of Freedom Estimates Ignoring Weights for the Dummy Regression Coefficient: BLACK

Model	Estimates		
	Linearization Standard Error $\sqrt{v_0}$	Corrected Standard Error $\sqrt{v_u}$ (with $T = \sigma^2 I_M$)	Effective Degrees of Freedom (with $T = \sigma^2 I_M$)
Standard Approach			
1	2.58	2.60	17.17
2	2.83	2.84	20.10
3	2.82	2.86	20.57
Ignoring the Stratification			
1	2.68	2.71	30.01
2	2.86	2.91	35.98
3	2.88	2.93	36.72
Ignoring the Stratification and the Clustering			
1	2.38	2.38	148.98
2	2.60	2.61	201.65
3	2.66	2.69	204.42

Model 1: $FAT = \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{WHITE} + \text{ERROR}$
 Model 2: $FAT = \beta_0 + \beta_1 \text{BLACK} + \text{ERROR}$
 Model 3: $FAT = \beta_0 + \beta_1 \text{BLACK} +$
 DUMMIES FOR DAY-OF-THE-WEEK
 and MONTH OF INTAKE + ERROR

Tables 3 and 4 mirror Tables 1 and 2 except that the single dummy of interest is women aged 65 to 69, while the second dummy is women aged 30 to 64. Since the two women with very large weights are under age 65, the gains in estimated degrees of freedom

resulting from removing the sample weights are quite small.

Sampled women aged 65 to 69 are in more PSU's than black women. As a result, their estimated effective degrees of freedom are higher than the corresponding estimates for black women when weights are ignored but clustering is *not*. By contrast, when stratification and clustering *is* ignored, the reverse is true. This is because there are more black women in the sample than women aged 65 to 69.

Surprisingly, the weights have less of an impact on the standard error estimates for the black women than for the older women. This is especially true for Model 3 where under the standard approach to variance estimation, the weighted standard error estimate for the black coefficient is less than the unweighted estimate!

Table 3. Standard Error and Effective Degrees of Freedom Estimates for the Dummy Regression Coefficient: AGE_65_TO_69

Model	Estimates		
	Linearization Standard Error $\sqrt{v_0}$	Corrected Standard Error $\sqrt{v_u}$ (with $T = \sigma^2 I_M$)	Effective Degrees of Freedom (with $T = \sigma^2 I_M$)
Standard Approach			
1	2.56	2.58	21.75
2	2.83	2.85	23.14
3	2.98	3.03	23.85
Ignoring the Stratification			
1	2.67	2.69	31.68
2	2.98	3.01	34.57
3	3.06	3.10	36.06
Ignoring the Stratification and the Clustering			
1	2.91	2.92	74.34
2	3.14	3.15	97.00
3	3.16	3.19	103.49

Model 1: $FAT = \beta_0 + \beta_1 \text{AGE}_65_TO_69 + \beta_2 \text{AGE}_30_TO_64 + \text{ERROR}$
 Model 2: $FAT = \beta_0 + \beta_1 \text{AGE}_65_TO_69 + \text{ERROR}$
 Model 3: $FAT = \beta_0 + \beta_1 \text{AGE}_65_TO_69 +$
 DUMMIES FOR DAY-OF-THE-WEEK
 and MONTH OF INTAKE + ERROR

This may reveal more about the instability of the weighted standard error estimates than about the true impact of the weights on the variances of the estimated coefficients. Nevertheless, we see by the example of black women that weight outliers can have a much

greater impact on inference than would be observable from looking at standard error estimates alone.

In addition to their grave effect on effective degrees of freedom calculations, the two weight outliers among the black women also appear to have an unusual impact on the bias of the linearization variance estimator. Only in Table 1 is the bias adjustment for the linearization standard error ever more than two percent. Even then, the estimated biases in Table 1 have a much smaller impact on confidence interval construction than the corresponding effective degrees of freedom calculations.

Table 4. Standard Error and Effective Degrees of Freedom Estimates Ignoring Weights for the Dummy Regression Coefficient: AGE_65_TO_69

Model	Estimates		
	Linearization Standard Error $\sqrt{v_u}$	Corrected Standard Error $\sqrt{v_u}$ (with $T = \sigma^2 I_M$)	Effective Degrees of Freedom (with $T = \sigma^2 I_M$)
Standard Approach			
1	2.03	2.04	23.44
2	2.22	2.23	24.89
3	2.33	2.35	25.57
Ignoring the Stratification			
1	2.08	2.09	38.91
2	2.31	2.33	42.39
3	2.40	2.43	43.67
Ignoring the Stratification and the Clustering			
1	2.38	2.38	117.99
2	2.60	2.61	149.59
3	2.66	2.69	154.87

Model 1: $FAT = \beta_1 AGE_{65_TO_69} + \beta_2 AGE_{30_TO_64} + ERROR$
 Model 2: $FAT = \beta_0 + \beta_1 AGE_{65_TO_69} + ERROR$
 Model 3: $FAT = \beta_0 + \beta_1 AGE_{65_TO_69} + DUMMIES\ FOR\ DAY-OF-THE-WEEK\ and\ MONTH\ OF\ INTAKE + ERROR$

6. DISCUSSION

In this paper we provided a mild generalization of the extended linear model framework in Kott (1991a) to cover regression with instrumental variables. Although all the theoretical results assumed that sampling was conducted without nonresponse or coverage error, it is a trivial matter to extend the

results. All that is needed is the existence of a known diagonal matrix of weights, W , with a zero in each diagonal corresponding to an element not in the respondent that also satisfies $\text{plim}_{M \rightarrow \infty} \mathbf{1}_M' W t / m = (p) \lim_{M \rightarrow \infty} \mathbf{1}_M' t / M$ for all relevant vectors t (such as $t = z, t = \epsilon, t = (t_1, \dots, t_N)'$ where $t_i = X_{(k)_i} z_i$, and so forth).

The argument repeated here that using survey weights provides protection against the possibility of model misspecification has an echo in the econometric literature. White (1980a) proposes using OLS rather than weighted least squares to protect against model misspecification of $E(y)$. In White's context, there are no survey weights. The weights he eschews are proportional to the inverses of the element variances (i.e., $w_i \propto 1/\text{Var}(\epsilon_i)$). Such weights are often used to increase the efficiency of linear regression estimators. Like a survey statistician, White is willing to sacrifice efficiency for robustness.

White (1980b) also proposes estimating the variance of a regression coefficient vector assuming that the element errors (the ϵ_i) are uncorrelated but that their variances are unknown. SAS has programmed this variance estimator in the ACOV option of PROC REG. White (1984) later generalized his result to cover more complex error structures. Under condition (v), this variance estimator is asymptotically equivalent to the one in equation (6).

We have seen that it is possible to correct for the bias in the linearization variance estimator (the bias in White's variance estimator could be removed similarly). To do so, we need to assume an exact model for the element error structure up to a constant. That assumption is precisely what we sought to avoid with the linearization variance estimator in the first place. Nevertheless, something has been gained: the resulting variance estimator has the same asymptotically small relative bias as the linearization variance estimator for the broad class of error structures satisfying condition (v).

We have also seen that at least for one particular data set the bias of a linearization variance estimator is less of a problem than its effective degrees of freedom. A method for estimating the effective degrees of freedom was proposed (equation (8)) but the method has three obvious drawbacks: (1) it relies on assumptions about the element error structure; (2) it relies on a normality assumption (of the ξ_{hj}); and (3) it only is asymptotically accurate even when all stipulated assumptions hold.

A sensitivity analysis with alternative assumptions about the element error structure is an obvious way to deal with drawback (1). In a similar manner, one

could do sensitivity analyses on a modified version of equation (7) with different values for $E(\xi_{hj}^4)/[E(\xi_{hj}^2)]$ to handle drawback (2). Even then, one needs to realize that the skewness of the ξ_{hj} , if any, can have a pronounced impact on one-sided hypothesis tests, an impact that is not addressed by effective degrees of freedom calculations.

More study may be needed to fully understand the repercussion of drawback (3). We can see from Tables 2 and 4 that for the unweighted domain means assuming no clustering and stratification, equation (8) produced estimated effective degrees of freedom that were biased upward by nearly one degree. For the black women in Table 2, 148.98 effective degrees of freedom were estimated rather than 148 (149 women minus 1 parameter, the domain mean). Fortunately, when constructing hypothesis tests and confidence intervals based on a Student t distribution with more than 10 degrees of freedom, cut-off values are not very sensitive to modest changes in the exact number of degrees of freedom.

One can show that when equation (8) applies asymptotically for the linearization variance estimator, it also applies for the jackknife and versions of the balanced repeated replication variance estimator when the last are calculable (see Kott, 1996). Unfortunately, employing equation (8) often requires matrix calculations. One of the popular reasons for computing replication variance estimators is to avoid such calculations.

Matrices are not always needed, however. It is a simple matter to compute equation (8) for an estimator of a domain mean under the assumption of uncorrelated and homoscedastic element errors: v_{hj} is replaced by the sum of the w_i^2 within PSU hj and the domain in question. This is a useful calculation to perform before naively assuming that a domain mean estimate is asymptotically normal.

When the estimated effective degrees of freedom for a regression coefficient is small, say less than 30, inferences based on normality should be avoided. This is a price one pays for using robust, randomization-based techniques. Two-sided inferences based on Student's t may be employed instead, but given the theoretical limitations of equation (8) discussed above, even they must be conducted with appropriate caution.

REFERENCES

- Fuller, W. A. (1975). "Regression analysis for sample survey," *Sankhyā*, Series, C, 37, 117-132.
- Fuller, W. A. (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. (1989). "Assessing linearization variance estimators," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 201-206.
- Kott, P. S. (1991a). "A model-based look at linear regression with survey data," *American Statistician*, 107-112.
- Kott, P. S. (1991b). "Estimating a system of linear equations with survey data," *Survey Methodology*, 91-98.
- Kott, P. S. (1994). "A Hypothesis test of linear regression coefficients with survey data," *Survey Methodology*, 159-164.
- Kott, P. S. (1996). "A model-based evaluation of several well known variance estimators for the combined ratio estimator," available from the author upon request.
- Schnell, D., Park, H. J., and Fuller, W. A., (1988). *EV CARP*, Ames, IA: Statistical Laboratory, Iowa State University.
- 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.
- USDA, Agricultural Research Service (1996). *1994 Continuing Survey of Food Intakes by Individuals and 1994 Diet and Health Knowledge Survey*, CD-ROM, accession no. PB96-501010, Springfield, VA: National Technical Information Service.
- White, H. (1980a). "Using least squares to approximate unknown regression functions," *International Economic Review*, 21, 149-170.
- White, H. (1980b). "A heteroskedasticity-consistent covariance matrix estimator and a direct test of heteroskedasticity," *Econometrica*, 48, 817-838.
- White, H. (1984). *Asymptotic Theory for Econometricians*, Orlando: Academic Press.