WHAT DOES PERFORMING A LINEAR REGRESSION ON SURVEY DATA MEAN?

Phillip S. Kott
NASS/USDA; S-4801; Washington, DC 20250

1. Introduction

A social 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.

Many survey statisticians have a quite different view of linear regression. They are interested in describing characteristics of a finite population. To this end, ordinary least squares regression performed on multivariate data from the entire population would produce some useful summary statistics. In practice, however, it is too difficult to obtain information from the entire population and so data is obtained from a sample of observations (note: the term "observation" will be used to refer to any member of the population under study even though relevant values for nonsampled members are not actually observed).

The social scientist's view of linear regression as given above is called model-based, the survey statistician's view design-based (Hansen et al. (1983)). According to model-based theory, part of the multivariate data -- the dependent variable -- is itself a random variable generated by a stochastic model. In contrast, orthodox design-based theory holds that all the data are fixed; the only thing probabilistic is the selection process that randomly chooses some observations for the sample and not others. There is no model generating the data. There is only a useful way to summarize the covariation of multivariate values in the finite population.

There is an alternative school of thought in design-based theory we will call the Fuller (1975, 1984) school. It holds that there is an underlying model generating the data, but that the analyst knows very little about it. In fact, the relationship among the variables may not even be linear. Linear regression is simply a means of summarizing in linear fashion a relationship among the multivariate values generated by the model.

There are several software packages that perform linear regressions and estimate variances in accordance with the Fuller school of design-based theory, which is more palatable to social scientists than the orthodox design-based approach. Two popular ones are SURREGR (Holt (1977)) and PC CARP (Fuller et al. (1986)).

This paper contrasts the three approaches to linear regression. It then shows how Fuller school procedures can offer protection against certain types of model failure from a model-based point of view. An illustrative example follows. A test for comparing the results of ordinary least squares and weighted regression is proposed.

2. The Standard Linear Model and the Sample

Suppose the multivariate values of a population of M observations can be fit by the linear model:

$$y = X\beta + \epsilon, \qquad (1)$$

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

If one knew y and X, then the best linear unbiased estimator of β would be the ordinary least squares (OLS) estimator

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

Unfortunately, y and X-values are only known for a sample of m observations which has been selected at random in a manner that is assumed to be independent of ϵ.

The best linear unbiased estimator of β given the information at hand is

$$b_{OLS} = (X'SX)^{-1}(X'Sy), \qquad (3)$$

where S is an $M \times M$ diagonal matrix of 0's and 1's. The ith diagonal of S is 1 if and only if the ith unit of the population is in the sample.

The variance of b_{OLS} (a variance-covariance matrix) is $\sigma^2(X'SX)^{-1}$. An unbiased estimator for this variance can be determined by estimating σ^2 in the above expression by
$s^2 = (y - Xb_{OLS})'S(y - Xb_{OLS})/(m - K)$.

3. The Design-Based Approaches

In the orthodox design-based approach to regression, there is no underlying linear model. The goal of linear regression is not to estimate β in equation (1); rather, it is to estimate B in equation (2) based on a randomly selected sample of m observations.

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

For many sampling designs the weighted regression estimator,

$$b_W = (X'WX)^{-1}(X'Wy), \qquad (4)$$

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

Fuller (1975) points out that b_W is generally a consistent estimator of $B^* = Q^{-1}R$, where $Q = \lim_{M->\infty}(X'X)/M$ and $R = \lim_{M->\infty}(X'y)/M$ when Q^{-1} and R exist and b_W is a consistent estimator of B. Often B is referred to as the finite population regression parameter, while B^* is the infinite population regression parameter.

What we have called the Fuller school of linear regression assumes the existence of a model generating the finite population data. It does not assume very much about the nature of that model, however, only that Q^{-1} and R exist. This school of thought employs the laws of probability in the same way as the orthodox design-based school does: through the sample selection process exclusively.

It should be noted that the model-based estimator, b_{OLS}, equals the design-based estimator, b_W, when W = S; that is, when all the sampled observations have equal probabilities of selection. It should also be noted that if the model in equation (1) holds, then the infinite population regression parameter, B^*, will equal the model regression parameter, β.

4. Design Mean Squared Error Estimation

In order to estimate the mean squared error of b_W as an estimator of either B or B^* under the sampling design, we need to know more about the design.

Suppose the population of M observations is divided into H strata (H may equal 1). Suppose further that there are $n_h \geq 2$ distinct primary sampling units (which may involve clusters of the actual observations) selected from stratum h. Ultimately, m_{hj} (which may also equal 1) observations are selected for the sample from primary sampling unit (PSU) hj. This broad framework allows for multistage random sampling with (perhaps) unequal selection probabilities at each stage. For simplicity, however, we exclude from consideration samples where some PSU has been selected more than once in the first sampling stage.

Without loss of generality, b_W can be rewritten as $b_W = Cy^*$, where y^* is an m-vector containing only those members of y that correspond to sampled observations. Let r^* be the analogously defined vector of residuals $(r = y - Xb_W)$.

For every sampled PSU hj, define D_{hj} as a m x m diagonal matrix of 1's and 0's such that the ith diagonal of D_{hj} is 1 if and only if the ith member of y^* corresponds to an observation in PSU hj. Finally, let $g_{hj} = CD_{hj}r^*$.

The linearization (or Taylor Series linearization or delta method) mean squared error estimator for b_W as an estimator of B^* is the matrix

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

This estimator is computed by the SURREGR software packages. PC CARP scales mse by {(m-1)/(m-K)}. Either way, the result is a consistent estimator of design mean squared error (in the Fuller school sense) as $n = \Sigma n_h$ grows arbitrarily large under mild conditions; see Shah et al. (1977) (note: orthodox design-based theory can require finite population correction terms which are unavailable in SURREGR and suppressible in PC CARP).

The Law of Large Numbers and the Central limit Theorem can often be invoked to test hypotheses of the form $HB^* = h_0$, where H is an r x k matrix and $r \leq k$. Under the null hypothesis,

$$T^2 = (Hb_W - h_0)'(H\{\text{mse}\}H')^{-1}(Hb_W - h_0) \qquad (6)$$

has an asymptotic chi-squared distribution with r degrees of freedom. When n - H - k is not large, a common ad hoc alternative to T^2 is $F = T^2/r$ which is assumed to have an F distribution with r and either n - H - K (SURREGR) or n - H (PC CARP) degrees of freedom.

5. The Extended Linear Model

In this section we will see that the use of b_W from equation (4) and mse from equation (5) can be justified in a purely model-based context. This is done by extending the linear model in equation (1) to allow for the possible existence of missing regressors and the likelihood that $\text{Var}(\epsilon)$ is much more complicated than $\sigma^2 I_M$.

Suppose the multivariate values of the population of M observations can be fit by the linear model:

$$y = X\beta + z + \epsilon, \qquad (7)$$

where y, X, β and ϵ are as before except that Var(ϵ) need not equal $\sigma^2 I_M$. The new vector z -- the putative missing regressor -- satisfies $\lim_{M\to\infty} X'z/M = 0$. It is a composite of all the regressors in a fully specified model for y that are otherwise missing from equation (7) and whose joint effect on y can not be captured within $X\beta$.

Under mild conditions, b_W is nearly (i.e., asymptotically) unbiased under the model (as n grows large), but the same can not be said for b_{OLS} unless $\lim_{M\to\infty} X'Pz/m = 0$, which in practical terms means that the probabilities of selection are unrelated to the missing regressors (proofs of these assertions are in Kott, 1991). Morevover, mse from equation (5) is a nearly unbiased estimator of the model mean squared error of b_W under many sampling designs and variance matrices for ϵ when $z \equiv 0$ and is reasonable when $z \not\equiv 0$ (see the appendix). The only restriction on Var(ϵ) is that $E(\epsilon_i \epsilon_{i'})$ be zero when i and i' are sampled observations from different PSU's and bounded otherwise. This is a very mild restric-tion since any covariation among observations across PSU's should, in principle, be captured by X or z.

The problem with b_W and mse from a model-based point of view is that they are not very efficient. For example, when z in equation (7) is identically zero and Var(ϵ) = $\sigma^2 I_M$, the variance of b_{OLS} will be less than that of b_W.

Note that even if Var(ϵ) $\not\equiv \sigma^2 I_M$, b_{OLS} is unbiased when $z \equiv 0$. Moreover, b_{OLS} may still be more efficient than b_W. With the g_{hj} in equation (5) appropriately redefined, mse could serve as an estimator of the variance of b_{OLS} under a fairly general specification for Var(ϵ). More efficient and also nearly unbiased (see the appendix or Kott, 1991) is

$$mse' = \frac{n}{n-1} \sum_{h=1}^{H} \sum_{j=1}^{n_h} g_{hj} g_{hj}', \qquad (8)$$

which equals mse when H = 1. It is a simple matter to get SURREGR and PC CARP to produce b_{OLS} and either mse' (SURREGR) or {(m-1)/(m-K)}mse' (PC CARP).

Although mse' (and mse for that matter) is an estimator for the variance of the estimated regression coefficient when $z \equiv 0$, we retain the "mse" notation for convenience.

Whether b_W or b_{OLS} is calculated, the test statistic in equation (6) can be employed (with b_{OLS} replacing b_W and perhaps mse' replacing mse as appropriate) to test hypotheses of the form $H\beta = h_0$.

6. An Example

Consider the folowing example synthesized from data from the National Agricultural Statistics Service's June 1989 Agricultural Survey. In a particular state, 17 primary sampling units have been selected from among 4 strata. These PSU's were then subsampled yielding a total sample of 252 farms. Although the sample was random, not all farms had the same probability of selection.

We are interested in estimating the parameters, β_1 and β_2, of the following equation:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + z_i + \epsilon_i, \qquad (9)$$

where i denotes a farm, y_i is farm i's planted corn to cropland ratio when i's cropland is positive, zero otherwise; x_{1i} is 1 if farm i has positive cropland, zero otherwise; and x_{2i} is farm i's cropland divided by 10,000.

Dropping all sampled farms with zero cropland from the regression equation will have no effect on the calculated values $b_{1\cdot W}$ and $b_{2\cdot W}$ (or $b_{1\cdot OLS}$ and $b_{2\cdot OLS}$) It would, however, affect mse (and mse') if none of the subsampled farms from a particular PSU had cropland. Although this phenomenon doesn't occur here, it does raise an issue worthy of a brief digression.

Sometimes a social scientist needs to perform a regression on a subset of a sample. In those circumstances, one may need to worry about the impact on mse when no member of the subset comes from a particular PSU. This problem can be avoided by treating all the originally sampled observations as if they were in the regression data set. Those observations not in the subset under study could be assigned y and x-values equal to 0.

The results of performing both OLS and weighted regression on the data in our example are displayed in Table 1. The table contains the square roots of mse and mse'. Also displayed is the square root of something denoted mse_0; this is the estimated mean squared error assuming that $z \equiv 0$ and that there is no correlation across observations within PSU's. Operationally, mse_0 is simply mse calculated as if there were 252 PSU's. The ACOV option of PROC REG in SAS (1985) will approximately yield this number (the value from ACOV needs to be multiplied by m/(m-1) for strict equality).

The ratio of mse/mse_0 is a measure of the effect of correlated errors within PSU's on the mean squared error of an estimated regression coefficent. This ratio will be greater than 1 when there is such a cluster effect. Similarly, the ratio mse/mse' is a measure of the effect of stratification on the mean squared error of an estimated regression coefficent. This ratio should be less than 1 when there is such a stratification effect (see the appendix). There can be cluster effects even when $z \equiv 0$, while there are stratification effects only when z_i values vary across strata. From Table 1, we can see there is generally much more pronounced cluster effects than stratification effects (if any).

7. A Test

Table 1 reveals that the OLS regression coefficients are more efficient (i.e, have smaller mse and mse' values) than the weighted regression coefficients. It remains to test whether these two sets of coefficients are really estimating the same thing. If that is the case, then the OLS estimates are clearly superior.

One general way to test whether b_{OLS} and b_W are estimating the same parameter vector, β, is to replace y in equation (4) by $y^e = (y', y')'$, X by

$$X^e = \begin{bmatrix} X & X \\ X & 0 \end{bmatrix},$$

and W by

$$W^e = \begin{bmatrix} W & 0 \\ 0 & S \end{bmatrix}.$$

The resulting estimator is $b^e{}_W = (b_{OLS}', d')'$, where $d = b_W - b_{OLS}$. Calculating mse^e is done in a manner analogous to mse in equation (5). Note that in calculating mse^e the elements of y^{e*} correspond to observations coming from the same number of PSU's (and strata) as do the elements of its analogue, y^*.

The test statistic in equation (6) can be invoked to test whether d is significantly different from 0 (with $b^e{}_W$ replacing b_W and mse_e replacing mse). This was done for the data set examined in the previous section. The resultant value for T^2 was 5.07. Observe that if T^2 is assumed to have a chi-squared distribution with two degrees of freedom, we would not reject the null hypothesis (that b_{OLS} and b_W are estimating the same thing) at the .05 significance level, although we would at the .1 level. Alternatively, assuming $T^2/2$ has an F distribution with 2 and 13 (17 PSU's minus 4 strata) degrees of freedom, the null hypothesis would not be rejected even at the .1 level.

This is not the end of the story however. If one's primary concern is robustness to the possible existence of a z vector related to the sampling weights rather than the efficiency of the estimated regression coefficients, then the fact that the test statistic exceeds its expected value under the null hypothesis (2 -- if T^2 is chi-squared) would be reason enough to prefer b_W over b_{OLS}.

Fuller (1984, eq. 17) proposed a different test for determining whether the difference between b_W and b_{OLS} is significant. His test assumes that the errors are independent and identically distributed across observations which is clearly not the case in our example.

Table 1-Estimated regression coefficients and root mean squared error estimates

Est. Reg. Coef.	Estimate	$\sqrt{mse}$	$\sqrt{mse'}$	$\sqrt{mse_0}$
$b_{1 \cdot W}$	.3363	.0822	.0781	.0301
$b_{2 \cdot W}$	.8636	1.2389	1.3008	.4764
$b_{1 \cdot OLS}$	.4460	.0396	.0440	.0192
$b_{2 \cdot OLS}$	-.8791	.4637	.4651	.1688

References

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," Sankhya, Ser. 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: Statistical Laboratory, Iowa State University.

Hansen M. H., Madow W. G., and Tepping, B. J. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," Journal of the American Statistical Association, 78, 776-793.

Holt M. M. (1977), SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data, Research Triangle Park: Research Triangle Institute.

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

SAS Institute (1985), SAS User's Guide: Statistics, Version 5 Edition. Cary, NC: SAS Institute.

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.

APPENDIX

The Near Unbiasedness of mse and mse' When $z \equiv 0$

Let $b_F = (X'TX)^{-1}X'Ty = C_T y^*$, where T can be either W or S. When $z \equiv 0$, b_T is an unbiased estimator for β with variance equal to $E(C_T \epsilon^* \epsilon^{*\prime} C_T')$. Since $Var(\epsilon^* \epsilon^{*\prime})$ is block diagonal, we can infer that

$$Var(b_T) = \Sigma_h \Sigma_j C_T D_{hj} \epsilon^* \epsilon^{*\prime} D_{hj} C_T'.$$

Let $g_{hj} = C_T D_{hj} r^*$, where $r^* = y^* - X^* b_T$. If n is large enough for ϵ^* and r^* to be nearly equal, then it is a simple matter to show that $E(mse') = E(\Sigma\Sigma\, g_{hj} g_{hj}') = E(\Sigma\Sigma\, C_T D_{hj} r^* r^{*\prime} D_{hj} C_T') \approx E(\Sigma\Sigma\, C_T D_{hj} \epsilon^* \epsilon^{*\prime} D_{hj} C_T') = Var(b_T)$. With similar reasoning, E(mse) can be shown to nearly equal $Var(b_T)$.

The Reasonableness of mse When $z \not\equiv 0$

When $z \not\equiv 0$, both b_W and b_{OLS} can be biased as estimators for β, but the latter is nearly unbiased under many sampling designs. Observe that the mean squared error of b_W is

$$
\begin{aligned}
MSE(b_W) &= \\
Var(b_W) &+ [Bias(b_W)]^2 \\
&= C_W Var(\epsilon\epsilon') C_W + [\sum_{h=1}^{H} \sum_{j=1}^{n_h} \tilde{f}_{hj}]^2 \\
&\approx C_W Var(\epsilon\epsilon') C_W + [\sum_{h=1}^{H} \sum_{j=1}^{n_h} f_{hj}]^2 \\
&= C_W Var(\epsilon\epsilon') C_W + [\sum_{h=1}^{H} (f_h - F_h)]^2, \qquad (A1)
\end{aligned}
$$

where q^2 denotes qq',

$$\tilde{f}_{hj} = (X'WX)^{-1}X'WD_{hj}[1 - (X'WX)^{-1}X'W]z,$$

$$f_{hj} = (M/m)(X'X)^{-1}X'WD_{hj}z, \quad f_h = \Sigma_j f_{hj},$$

and F_h is the limit of the <u>design</u> expectation of f_h as N_h (the number of PSU's in h) grows arbitrarily large (note that $\Sigma F_h = 0$ since $\lim_{M\to\infty} (X'X)^{-1}X'z = 0$).

It is a simple matter to show that

$$E(mse) \approx C_W Var(\epsilon\epsilon') C_W + \sum_{h=1}^{H} \frac{n_h}{n_h - 1} [\sum_{j=1}^{n_h} f_{hj}^2 - \frac{f_h^2}{n_h}]. \qquad (A2)$$

When all the N_h are assumed to be arbitrarily large, the distinction between with and without replacement sampling of PSU's is lost and the <u>design</u> expectations of the right hand sides of (A1) and (A2) coincide.

Similar to (A2) is

$$E(mse') \approx C_W Var(\epsilon\epsilon') C_W + \frac{n}{n-1} \sum_{h=1}^{H} \sum_{j=1}^{n_h} f_{hj}^2. \qquad (A3)$$

Clearly, a diagonal element of E(mse') will exceed the corresponding element of E(mse) when the corresponding diagonal of Σf_h^2 exceeds that of $\Sigma (\Sigma f_{hj}^2)$. This tends to be the case when the appropriate elements of the F_h are not all identically zero; that is, when the effect of the putative missing regressor, z, varies across strata.