What Does Performing Linear Regression on Sample Survey Data Mean?

Phillip S. Kott

Abstract. *Most economists understand linear regression as the estimation of the parameters of a linear model. There are two other ways of interpreting the results of linear regression, however, and most software packages designed specifically to handle data from complex sample surveys (for example, SURREGR and PC CARP) assume one of these interpretations. This article contrasts the conventional model-based theory of linear regression to the design-based theories underlying survey-sampling software. The article demonstrates how procedures from design-based regression theory can be justified and exploited in a linear model framework. Proposed is a test for comparing the results of ordinary least squares and weighted regression.*

Keywords. *Design-based, model-based, random sample, mean-squared error*

An economist 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, though, have a 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 can 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. (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 economist's view of linear regression as given above is called "model-based," the survey statistician's view "design-based" (*4*).[1]

According to model-based theory, part of the multivariate data—the dependent variable—is itself a random variable generated by a stochastic model. Orthodox design-based theory, in contrast, 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, 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 that we will call the "Fuller School" (*1*, *2*). This theory says that although there is indeed an underlying model generating the data, the analyst knows little about this model. 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.

Several software packages perform linear regressions and estimate variances in accordance with the Fuller School, which is more palatable to economists than the orthodox design-based approach. Two popular packages are SURREGR (*5*) and PC CARP (*3*).

The Standard Linear Model and the Sample

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

$$y = X\beta + \epsilon, \tag{1}$$

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 $Var(\epsilon) = E(\epsilon\epsilon') = \sigma^2 I_M$.

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

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

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

The best (minimum variance) linear unbiased estimator of β, given the sample, is

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

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

---

Kott is special assistant for economic survey methods in the Office of the Director, Bureau of the Census, and was senior mathematical statistician with the Survey Research Branch, National Agricultural Statistics Service.

[1]Italicized numbers in parentheses cite sources listed in the References at the end of this article.

equation 3 allows only those rows of X and elements of y containing information from sampled observations to be captured in b_{OLS}

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 $s^2 = (y - Xb_{OLS})'S\,(y - Xb_{OLS})/(m - K)$

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 an $M \times M$ diagonal matrix, the ith diagonal of which is the probability that 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,

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

is a design-consistent estimator of B in equation 2 That is, as m (and M) grows arbitrarily large, b_W - B has a probability limit of zero with respect to the probability space generated by the sampling mechanism

Fuller (*1*) points out that b_W is generally a consistent estimator of $B^* = Q^{-1}R$, where $Q = \lim_{M\to\infty} (X'X)/M$ and $R = \lim_{M\to\infty} (X'y)/M$ 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, but not assuming very much about the nature of that model, only that Q^{-1} and R exist This theory employs the laws of probability in the same way as the orthodox design-based school does exclusively through the sample selection process

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) If the model in equation 1 holds, then the infinite population regression parameter, B^*, will equal the model regression parameter, β

Design Mean-Squared Error Estimation

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 L strata (L may equal 1) And, suppose 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 the 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 and C is the m corresponding columns of $(X'WX)^{-1}X'W$ Let r^* be the vector of residuals analogous to y^* (note $r = y - Xb_W$)

For every sampled PSU hj, define D_{hj} as an $m \times m$ diagonal matrix of 1's and zeroes such that the ith diagonal of D_{hj} is 1 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

$$\text{mse} = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \Big[\sum_{j=1}^{n_h} g_{hj}g_{hj}' - \frac{1}{n_h} \big(\sum_{j=1}^{n_h} g_{hj}\big) \big(\sum_{j=1}^{n_h} g_{hj}\big)'\Big] \tag{5}$$

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 (*8*) (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 $r \times K$ matrix and $r \leq K$ Under the null hypothesis,

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

has an asymptotic chi-squared distribution with r degrees of freedom When n - L - 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 – L –K (SURREGR) or n – L (PC CARP) degrees of freedom

The Extended Linear Model

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 Var(ϵ) is much more complicated than $\sigma^2 I_M$ The proofs for the assertion made in this section and other technical details are in (*6*)

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

$$y = X\beta + z + \epsilon, \tag{7}$$

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

Under mild conditions, b_W is nearly (that is, asymptotically) unbiased under the model in equation 7 (as n grows large) The same cannot be said for b_{OLS} unless $\lim_{M \to \infty} \bar{X}'Pz/m = 0$, which in practical terms means that the probabilities of selection are unrelated to the missing regressors

The expression in 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 The only restriction on the latter is that $E(\epsilon_i \epsilon_{i'})$ be zero when i and i' are sampled observations from different PSU's and bounded otherwise This restriction is very mild 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

Even if Var(ϵ) $\neq \sigma^2 I_M$, 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 is the matrix,

$$\text{mse}' = \frac{n}{n-1} \sum_{h=1}^{L} \sum_{j=1}^{n_h} g_{hj} g_{hj}', \tag{8}$$

which equals mse when L = 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 $H\beta = h_0$

An Example

Consider the following example synthesized from USDA data from the National Agricultural Statistics Service's June 1989 Agricultural Survey In a particular State, 17 primary sampling units were 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

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

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_{1W} and b_{2W} (or b_{1OLS} and b_{2OLS}) It would, however, affect mse (and mse') if none of the subsampled farms from a particular PSU had cropland Although this phenomenon does not occur here, it does raise an issue worthy of a brief digression

Sometimes an economist 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 estimated root mean-squared errors computed from the appropriate diagonal elements of mse and mse' Also displayed is $\sqrt{\text{mse}_0}$ the estimated coefficient root mean-squared error assuming that $z \equiv 0$ and that there is no correlation across observations within PSU's The variance matrix mse_0 is simply mse' calculated as if there were 252 PSU's The ACOV option of PROC REG in the popular programming language SAS (*7*) used along with a weight statement 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 coefficient. 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 coefficient. This ratio should be less than 1 when there is such a stratification effect.

There can be cluster effects even when $z \equiv 0$, while there are stratification effects only when z_i values vary across strata. We can see from table 1 that there are generally much more pronounced cluster effects than stratification effects (if any).

A Test

Table 1 reveals that the OLS regression coefficients are more efficient (that is, 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

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

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. 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 zero (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. If T^2 is assumed to have a chi-squared distribution with two degrees of freedom, the null hypothesis was not rejected (that b_{OLS} and b_W are estimating the same thing) at the 0.05 significance level but would be rejected at the 0.1 level. 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 0.1 level.

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 (*2*, p. 106, equation 17) proposed a different test for determining whether the difference between b_W and b_{OLS} is significant. His test assumed that the errors were 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

Estimated regression coefficient	Estimate	$\sqrt{mse}$	$\sqrt{mse'}$	$\sqrt{mse_0}$
b_{1W}	0.3363	0.0822	0.0781	0.0301
b_{2W}	.8636	1.2389	1.3008	.4764
b_{1OLS}	.4460	.0396	.0440	.0192
b_{2OLS}	–.8791	.4637	.4651	.1688

References

1. Fuller, W. A. "Regression Analysis for Sample Survey," *Sankhya*. Ser. C, 37, 1975, pp. 117-32.

2. ______. "Least Squares and Related Analyses for Complex Survey Designs," *Survey Methodology*. Vol. 10, 1984, pp. 97-118.

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

4. Hansen, M. H., W. G. Madow, and B. J. Tepping. "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," *Journal of the American Statistical Association*. Vol. 78, 1983, pp. 776-93.

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

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

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

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