OUTLIERS IN REGRESSION ANALYSIS WHEN MEASUREMENT ERROR IS LARGE

Eugene P. Ericksen, Institute for Survey Research, Temple University

I. Introduction and Description of Regression--Sample Data Technique

The assumption of linearity is critical for the accurate estimation of coefficients in linear least squares regression. While observed nonlinearities commonly arise because relationships are truly curved, they can also occur when linearly related variables are measured with error, and the errors are consistently related to the true values. It is also possible that a few grossly outlying observations can give the appearance of nonlinearity, particularly if they are located near the extremes of the distribution. As Hogg (1974), in a recent review of the issue, quotes Huber, "Just a single grossly outlying observation may spoil the least squares estimate, and moreover, outliers are much harder to spot in the regression than in the simple location case." This point is all the more relevant when the outliers are the result of measurement error. Because the estimation of curvilinear relationships is difficult, and even misleading when measurement error obscures a truly linear relationship, statisticians have advocated other methods which reduce or eliminate the impact of nonlinearities and outliers. Three such approaches are (1) the transformation of data through logarithmic, arc sine, or other rules, (2) the elimination of or reduction in the weights assigned to outliers, and (3) the use of Stein-James estimators. We have used these to attempt to reduce the effects of nonlinearities and measurement error in applying the regression-sample data method (Ericksen, 1973, 1974) to estimate postcensal population growth.

The regression-sample data method proceeds as follows:

a. Sample estimates are computed for each of the primary sampling units (psus) included in an available sample, in this case the Census Bureau's Current Population Survey. The dependent variable is the ratio of the 1970 CPS sample estimate for the psu to the recorded 1960 Census population.

b. Symptomatic indicators of population growth, assumed to be measured without error, are compiled, and used as independent variables in regression. Examples of these are the 1970/1960 ratios of births, deaths, and school enrollment and alternative estimates of population growth computed by the ratio-correlation technique and Census Method II (for discussion and references describing these procedures, see U.S. Bureau of the Census, 1973).

c. Regression equations are then computed and used to calculate population estimates for sample psus and for all counties where symptomatic data are available.

In 1970 these regression estimates were found to be more accurate than estimates computed by any other procedure then in use. However, they were not as accurate as they would have been in the absence of the within-psu sampling error. The mean squared error of these regression estimates is expressed by the following:

$$MSE = \frac{(n - p - 1)\sigma_u^2}{n} + \frac{(p + 1)\sigma_v^2}{n} \qquad (1)$$

where n = the weighted sum of psus in the sample,

p = the number of independent variables,

σ_v^2 = the within-psu error, or the mean squared error of the sample estimates, and

σ_u^2 = the residual between-psu error that would be obtained from least squares regression if there were no within-psu error.

The presence of the within-psu error at least partially offsets the gains made by adding symptomatic indicators. If sample estimates with large errors could be removed or reduced in importance, the within-psu error would be reduced and more efficient use of available symptomatic information could be made.

To illustrate the effect of the within-psu error we can refer to Table 1 to compare the results obtained when CPS sample estimates were used as the dependent variable (Series A) to those obtained when 1970/1960 Census ratios were used (Series B) and the within-psu error thus eliminated. The errors for Series A were larger. Furthermore, where gains could be made in Series B by adding symptomatic indicators, these were largely offset in Series A by increases in the within-psu error.

Although not shown in Table 1, there was some indication that relationships were not linear throughout the distribution. Values at the extremes, particularly the positive extreme, tended to have larger errors. As an initial attempt to rectify this we computed a regression equation with all variables in logarithmic form. Unfortunately, the result, given as Series C in Table 1, was that a slight increase in error was obtained. It is possible that a different transformation would have given better results but the difficulty of selecting the best transformation in advance is troublesome. Moreover, it appears that a small to moderate number of outlying observations is causing most of the difficulty. A more promising strategy seems to be to eliminate or reduce the effects of these outliers.

2. Dealing with Individual Outliers

Although Hogg (1974) describes or refers to many strategies of dealing with outliers in regression which have produced good results on various sets of data, there does not appear to be a set of empirical rules to determine the best strategy in advance. In these strategies a preliminary estimate of the regression equation is computed, and cases where the regression estimate and observed value are greatly different are either removed from the sample or assigned a lower weight. The trick is to remove enough outliers so that errors are reduced without removing so many observations that the regression equation computed on the basis of remaining observations is biased. For our applied problem, the fact that the measurement error was large relative to the errors in regression lessened this difficulty.

To be specific, the mean squared error of the psu sample estimates, or the within-psu error, was found to be $\sigma_v^2 = .0243$ and the mean squared error of various regression estimates ranged from .0014 to .0016. Therefore a sample estimate which deviated grossly from a regression estimate was also likely to deviate grossly from the corresponding true, but unobserved, value. Rules were thus defined which eliminated or reduced the weights assigned to observations where the preliminary regression estimate was very different from the sample estimate.

The estimated mean squared difference between the regression and sample estimates is given by the formula:

Mean Squared Difference =

$$\frac{(n - p - 1)}{n}(\sigma_u^2 + \sigma_v^2) \qquad (2)$$

where the terms are defined as in equation (1). A preliminary regression equation was computed using the ratios of births, deaths, and school enrollment as symptomatic indicators and the value given by equation (2) was a mean squared difference of .0254 with the root mean squared difference thus about .16.

A gross outlier was defined as an observation where the difference between the sample and regression estimates was over (2)(.16) = .32 and a moderate outlier was an observation where this difference was between .16 and .32. Assuming normality of the within-psu sampling distributions we would have expected about five percent of the observations to be gross outliers if only random error had been present. In fact, as in any survey of human population, there were nonrandom sources of error such as differential nonresponse, clerical and coding error, records which could not be read by the computer and other errors in data compilation and processing. Somewhat more than five percent of the observations were gross outliers.

Each psu had previously been assigned a weight depending on the size of the stratum it represented and hence its sample size and presumably the accuracy of the sample estimate. For some rules, the weight was now set equal to zero and the psu was eliminated from the computation of the regression equation. In other cases the weight was divided by two. The various rules and results of these rules are presented in Tables 1 and 2. In all cases regression estimates and errors were computed for all sample psus, whether they were included in the computation of the regression equation or not. For example, three psus, each with a weight of 1.0 were eliminated under Rule D and the regression equation was computed on the basis of the remaining 386 psus. Regression estimates were computed for all 389 psus and the accuracy of these 389 estimates is reported in Table 1. At one extreme (Rule D) only the most grossly outlying observations were eliminated, while at the other extreme (Rule J) all gross and moderate outliers were eliminated. Rules D, E, and G eliminated some or all of the gross outliers while Rules F and H simply reduced the weights assigned to them. Rule I eliminated gross outliers and reduced the weight assigned to moderate outliers while Rule J eliminated both gross and moderate outliers. Rules D and J were formulated as extremes with the expectation that the best results would occur somewhere in between. As it happened, noticeable improvements were obtained for all rules except Rule J where the results were no worse than for Series A. The results can be summarized as follows:

a. Referring to Table 1, the best results were obtained when the gross outliers were eliminated. When the effects of the gross outliers were merely reduced the gains were not as great. Similarly, when the effects of the moderate outliers were reduced or eliminated, the gains were also smaller. Finally, the results of all these schemes except Rule J were improvements over using the raw data in Series A and in some cases the mean squared error was reduced by 15 to 20 percent.

b. For all except Rule J, clear improvements were obtained when four or five symptomatic indicators were used. Reduction of the within-psu error thus makes possible a more efficient use of available symptomatic information.

c. Referring to Table 2, we can see that the regression coefficients obtained using Rules D, E, and G were very similar to those obtained in Series B when no within-psu error was present. Similarly, the level of error was about the same in Series B, D, E, and G.

The regression-sample data estimates were thus improved by the elimination of gross outliers. We cannot give a procedure for selecting one best rule, but clear gains can be obtained from any one of several sensible appearing rules.

If the distributions of the sample means for individual psus are approximately normal, one would expect about five percent of the observations to be greater than twice the square root of the estimated within-psu variance. This in turn is only slightly less than the root mean squared difference between the regression and sample estimates. It therefore seems reasonable to formulate a rule based on the elimination of gross outliers as defined in this illustration. In cases where the measurement errors are not as large relative to the errors of regression a different rule might be more appropriate.

3. Using Stein-James Estimates

Our third approach was to compute Stein-James estimates of the sample psu observations and then to use these estimates as the dependent variable in regression. Lindley (1962), adding to the basic result of James and Stein (1961) showed that where one wished to estimate the parameters $\theta_1, \theta_2, \ldots, \theta_k$ where each θ_i is the mean of an independent normal variate X_i, where X_i was distributed with mean θ_i and a common variance σ_v^2, for $k \geq 4$, the estimator

$$\delta_i = \bar{X} + \left[1 - \frac{(k-3)\sigma_v^2}{S'}\right](X_i - \bar{X}), \qquad (3)$$

where $S' = \sum_{i=1}^{k}(X_i - \bar{X})^2$,

is uniformly better than the maximum likelihood estimator, which in our case is the set of sample means for the sample psus. Other descriptions and references explaining the method and its uses are given by Efron and Morris (1973, 1975). Where θ_i lie close together and σ_v^2 is large, significant gains in accuracy have been obtained (e.g., Efron and Morris, 1975). Translating the symbols into the terminology used in this paper, the θ_i represents the 1970/1960 Census ratios of population growth, the X_i represent the CPS sample estimates, σ_v^2 represents the within-psu error, as in equation (1), and k represents the number of primary sampling units. We found that we were able to obtain significant increases in the accuracy of the sample estimates using the Stein-James estimator. Unfortunately these were not translatable into improvements in the regression estimates.

In order to apply the Stein-James estimator, certain conditions are necessary. These are:

a. Each X_i is normally distributed. Given the within-psu sample sizes available, this was approximately true.

b. The value of σ_v^2 is the same for all X_i. This condition was met by restricting our attention only to 297 psus where the sample size, and hence the expected value of σ_v^2, were about the same.

c. The values of the θ_i had to be close. This was accomplished by grouping observations using the preliminary regression estimate based on births, deaths, and school enrollment used to spot outliers in the preceding discussion. Because the distribution of θ_i was positively skewed, more subgroups were needed at the top of the distribution.

The question of how many subgroups to use is critical. We tried successively subdividing the sample into 1, 2, 4, 10 & 27 subgroups of nearly equal size. As can be seen from equation (3), the Stein-James estimator has the effect of pulling values into the subgroup mean. Where most of the variation in a subgroup is sampling error, this has the effect of reducing or eliminating the within-psu error. Where the variance in actual values is larger, this pulls in the actual values as well, thus biasing the estimate toward the subgroup mean. For example, when only one subgroup was used, the variance of the distribution of Stein-James estimates was smaller than either the variance of the distribution of sample estimates or the distribution of actual 1970/1960 Census ratios. This is because the Stein-James estimates at the top end of the distribution were consistently lower than either the Census recorded values or the CPS sample estimates and the opposite was true at the bottom end of the distribution. The Stein-James estimates in the middle of the distribution tended not to have a consistent error in direction and, as shown in Table 3, were considerably more accurate than the CPS sample estimates. On the other hand, when the number of subgroups used was increased, the Stein-James estimates for the extreme groups, particularly at the top end of the distribution, became more accurate. Unfortunately, when the number of subgroups was large the errors of Stein-James estimates for psus in the middle range increased. This is presumably due to the effects of errors in the preliminary regression estimates. As shown below, the most efficient overall results were obtained by splitting the sample into four or ten subgroups.

Estimator	Mean Squared Error	Percent Reduction in Error
Original Sample Estimates	.0343	----
Stein-James, 1 Subgroup	.0136	60.3
Stein-James, 2 Subgroups	.0096	72.0
Stein-James, 4 Subgroups	.0077	77.6
Stein-James, 10 Subgroups	.0078	77.3
Stein-James, 27 Subgroups	.0101	70.6

As shown in Table 3, only the errors of the top decile of psus were minimized by sorting into 27 subgroups. In all other deciles, the mean

squared error was minimized by sorting the sample into four or ten subgroups. Clearly, the best results in regression could be obtained by using as the dependent variable the Stein-James estimates based on 27 subgroups for the top decile and the Stein-James estimates based on 4 or 10 groups for all remaining observations.

The seven distributions of Stein-James estimates indicated in Table 3 were finally used as dependent variables in regression with four or five symptomatic indicators. The accuracy of these regression estimates is indicated in Table 4. We can see that little or no gain was obtained except for distribution 27-04, i.e., where the Stein-James estimates based on 27 subgroups were used for the top decile and the Stein-James estimates based on 4 subgroups were used for all other psus. This particular distribution of Stein-James estimates was selected using the information on their accuracy obtained by comparison with Census data. It is doubtful if such an appropriate selection could be made in practice. It appears that if Stein-James estimates, while they clearly improve the accuracy of the sample estimates, are to be useful in regression analysis, an efficient algorithm for grouping observations will have to be developed. Otherwise, the procedure is too complicated to use for such moderate gains.

To summarize, the best results were obtained by first using a preliminary set of regression estimates to spot individual outliers and then eliminating or reducing the weights assigned to these outliers. This made possible the subsequent calculation of more accurate regression equations. We have not solved the problem of identifying the best rule for spotting and reducing the impact of these outliers. We have shown, however, that where the measurement error of the observations is large relative to the errors in regression, there is a large class of rules which lessen the effects of outliers and lead to more accurate estimation of regression equations. For our regression-sample data estimates, using these rules sharply reduced the within-psu component of the mean squared error and gains of up to 20 percent in the overall mean squared error were obtained. In fact, the results were nearly as good as if the within psu component of error had been removed completely. There are two important next steps. The first is to see whether similar gains could be obtained for other variables and/or other time periods. The second is to devise an estimator for the mean squared error when outliers are removed or reduced in weight.

BIBLIOGRAPHY

Efron, B. and Morris, C., "Combining Possibly Related Estimation Problems," Journal of the Royal Statistical Society, Series B, 35, 379-421, 1973.

Efron, B. and Morris, C., "Data Analysis Using Stein's Estimator and its Generalizations," Journal of the American Statistical Association, 70, 311-319, 1975.

Ericksen, E.P., "A Method for Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas," Demography, 10, 137-160, 1973.

Ericksen, E.P., "A Regression Method for Estimating Population Changes of Local Areas," Journal of the American Statistical Association, 69, 867-875, 1974.

Hogg, R.V., "Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory," Journal of the American Statistical Association, 69, 909-923, 1974.

James, W. and Stein, C., "Estimation With Quadratic Loss," in Proceedings of the Fourth Berkeley Symposium, University of California Press, Volume 1, 361-379, 1961.

Lindley, D.V., "Discussion" of C. Stein, "Confidence Sets for the Mean of a Multivariate Normal Distribution," Journal of the Royal Statistical Society, Series B, 24, 265-296, 1962.

U.S. Bureau of the Census, "Federal-State Cooperation Program for Local Population Estimates: Test Results--April, 1970," Current Population Reports, Series P-26, No. 21. Washington, D.C., Government Printing Office, 1973.

ACKNOWLEDGEMENT

The research upon which this paper is based was carried out in cooperation with the Bureau of the Census, in particular with the Research Center for Measurement Methods and with the Population Division who supplied the financial support and necessary data. In particular, discussions with Kirk Wolter and David Word have been helpful. I would also like to thank Harris Miller of Temple University without whose diligence in computer programming this research could not have been carried out. The findings, recommendations, and conclusions in this paper are the sole responsibility of the author and are not necessarily endorsed by the U.S. Government. The data in this paper are the result of tax-supported research and, as such, are not copyrightable. The data may be freely reprinted with the customary crediting of the source.

Table 1: Errors Obtained by Various Regression Equations for Estimates of Population Growth in 1960-70 for 389 Primary Sampling Units

Series[1] (Sum of Weights)	Number of Symptomatic Indicators[2]: 2 Mean Absolute Percent Error	2 Mean Squared Error ($\times 10^{-4}$)	3 Mean Absolute Percent Error	3 Mean Squared Error ($\times 10^{-4}$)	4 Mean Absolute Percent Error	4 Mean Squared Error ($\times 10^{-4}$)	5 Mean Absolute Percent Error	5 Mean Squared Error ($\times 10^{-4}$)
A (647)	2.81	14.00	2.97	15.59	2.91	14.68	2.89	14.44
B (647)	2.76	14.27	2.71	13.64	2.59	12.30	2.54	11.80
C (647)	3.25	17.14	3.34	17.99	3.31	17.47	3.35	17.89
D (644)	2.77	13.94	2.72	13.50	2.63	12.25	2.62	12.11
E (636)	2.75	13.83	2.69	13.26	2.58	12.06	2.53	11.51
F (633)	2.75	13.75	2.82	14.39	2.73	13.28	2.69	12.87
G (619)	2.76	14.00	2.73	13.78	2.61	12.51	2.56	12.04
H (584)	2.82	14.12	2.88	14.66	2.80	13.68	2.81	13.85
I (570)	2.81	14.37	2.78	14.05	2.66	12.81	2.63	12.58
J (521)	2.97	15.45	2.91	15.04	2.79	13.93	2.87	14.75

[1] Definitions of series. Dependent variable is:
A: Original sample estimates obtained from CPS.
B: Census recorded values.
C: Logarithms of CPS estimates (all other variables in logarithmic form).
D: Set $Z = (Y_o - \hat{Y})/\hat{Y}$ where Y_o = the CPS sample estimate and $\hat{Y}$ = the preliminary regression estimate. Eliminate if $Z > .64$.
E: Eliminate if $Z > .48$.
F: Divide weight in half if $Z > .32$.
G. Eliminate if $Z > .32$.
H. Divide weight in half if $Z > .16$.
I. Eliminate if $Z > .32$ and divide weight in half if $.16 < Z < .32$.
J. Eliminate if $Z > .16$.

[2] Independent variables used for:
2 Symptomatic indicators, school enrollment and ratio-correlation estimate.
3 Symptomatic indicators, births, school enrollment and ratio-correlation estimate.
4 Symptomatic indicators, births, deaths, school enrollment, and ratio-correlation estimate.
5 Symptomatic indicators, births, deaths, school enrollment, ratio-correlation estimate, and method II estimate.

Table 2: Regression Coefficients Obtained for Four Variable Equations

Series[1]	Constant	Births	Deaths	Enrollment	Ratio-Correlation	Coefficient of Determination, R^2
A	.058	-.097	+.045	+.214	+.745	.428
B	.004	+.094	+.111	+.229	+.573	.951
C	-.035	-.050	+.024	+.294	+.665	.404
D	.049	+.062	+.085	+.218	+.588	.439
E	.038	+.088	+.072	+.242	+.570	.491
F	.049	-.040	+.059	+.256	+.657	.482
G	.039	+.029	+.076	+.299	+.556	.552
H	.047	-.036	-.046	+.276	+.646	.534
I	.036	+.039	+.064	+.325	+.534	.632
J	.033	+.055	+.050	+.364	+.496	.770

[1] See Table 1 for definitions.

Table 3: Mean Squared Errors of Stein-James Estimates Using Various Grouping Strategies

Decile	Original CPS Sample Estimates	Number of Groups[1] 1	2	4	10	27	27-04	27-10	Size of Group
1	.0510	.0349	.0259	.0216	.0190	.0145	.0145	.0145	30
2	.0393	.0130	.0073	.0085	.0069	.0105	.0085	.0069	30
3	.0247	.0113	.0063	.0059	.0038	.0057	.0059	.0038	30
4	.0265	.0074	.0071	.0032	.0043	.0052	.0032	.0043	30
5	.0399	.0118	.0071	.0029	.0058	.0103	.0029	.0058	30
6	.0492	.0133	.0126	.0095	.0127	.0175	.0095	.0127	30
7	.0459	.0124	.0125	.0089	.0088	.0136	.0089	.0088	30
8	.0259	.0086	.0059	.0068	.0061	.0107	.0068	.0061	29
9	.0183	.0088	.0042	.0044	.0046	.0051	.0044	.0046	29
10	.0206	.0140	.0064	.0047	.0057	.0074	.0047	.0057	29
Total	.0343	.0136	.0096	.0077	.0078	.0101	.0070	.0073	

[1] This refers to the number of groups the psus were divided into before the Stein-James estimates were computed. Where the number of groups is given as 27-04, the top 30 observations were grouped as they were in the subdivision into 27 subgroups while remaining observations were given the values assigned when four subgroups were created. Similarly, where the number is given as 27-10, the top 30 were assigned the values obtained when there were 27 subgroups while remaining observations were assigned the values given when there were ten subgroups. The subgroups were defined on the basis of a preliminary set of regression estimates.

Table 4: Errors of Regression Estimates When Stein-James Estimates Are Used as Dependent Variable for 297 Psus

Number of Symptomatic Indicators[2]		Definition of Dependent Variable[1]: Stein-James Estimates 1	2	4	10	27	27-04	27-10	Original CPS Sample Estimates	1970/1960 Census Ratios
Four	Mean Absolute Percent Error	5.83	4.46	4.04	3.96	3.80	3.54	3.87	3.85	2.79
	Mean Squared Error x 10^{-4}	54.35	31.55	24.95	23.14	21.33	19.25	22.16	22.26	13.96
Five	Mean Absolute Percent Error	5.89	4.46	4.03	3.94	3.78	3.52	3.84	3.83	2.77
	Mean Squared Error x 10^{-4}	54.52	30.11	22.71	20.23	20.53	18.63	21.55	21.83	13.35

[1] Definitions of the Dependent Variables are the same as in Table 3.

[2] The symptomatic indicators used to compute the regression equations are the same as indicated in Table 1.