This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: Annals of Economic and Social Measurement, Volume 3, number 2

Volume Author/Editor: Sanford V. Berg, editor

Volume Publisher: NBER

Volume URL: http://www.nber.org/books/aesm74-2

Publication Date: April 1974

Chapter Title: Detecting Errors in Economic Survey Data: Multivariate vs Univariate Procedures

Chapter Author: Philip Musgrove

Chapter URL: http://www.nber.org/chapters/c10113

Chapter pages in book: (p. 333 - 345)

Annals of Economic and Social Measurement, 3/2, 1974

DETECTING ERRORS IN ECONOMIC SURVEY DATA: MULTIVARIATE VS. UNIVARIATE PROCEDURES

BY PHILIP MUSGROVE

Errors are sought in a large body of household survey data by using prior knowledge of relations among variables, rather than assumptions about the distribution of the errors. Provided the errors are confined to the dependent variable of a linear regression model, the residuals from the regression can be used to identify probably-erroneous observations. This test is compared, in efficiency and thoroughness, to a univariate test which detects only extremely high or low observations.

ACKNOWLEDGEMENTS

I am indebted to several colleagues, none of whom is to blame for the faults of the paper. The notion of the cost and benefit of an extreme value test derives, with some modification, from a proposal by Arturo Meyer for comparing a number of tests. The multivariate test was developed by Howard Howe; discussion of it with Denisard Alves was most helpful. The empirical results were assembled by Ximena Cheetham, Marily Luders, Jorge Lamas and Dana Nowicki. One of the computer programs used was written by Roberto Villaveces; the other was written by Marcia Mason. Robert Ferber advised on the tests developed and commented on some preliminary ideas for this essay. Howard Howe reviewed the first draft and offered several helpful criticisms. William Madow commented on the paper as it was first presented. Lester Taylor offered a suggestion for treating one of the principal problems noted.

1. INTRODUCTION

This paper considers how to detect errors in quantitative continuous variables of the kind obtained in survey data. "Errors" include the readily apparent extreme values together with misreported values which may lie near the center of a variable's distribution but still differ significantly from the true values. The procedures and results discussed derive from the experience of trying to correct errors in a large body of household budget data.

Both the purpose of the inquiry and the assumptions employed in it differ from those associated with many errors-in-variables problems. These differences are briefly described, with reference to some of the literature, in the bibliographic note at the end.

2. SOME CHARACTERISTICS OF TESTS FOR ERRORS

Suppose there are T observations of each of n variables. We define a *test* as any procedure for selecting from this matrix η observations of one variable. (n is the number of variables used in the test, which may be much less than the number available in the data.) If $n = 1$, the test is *univariate*; if $n > 1$, the test is

multivariate. Let v be the number of errors present in the variable being tested, and let $v(\eta)$ be the number of such errors detected in η observations.

We define the *efficiency* of a test as $\theta(\eta) = v(\eta)/\eta$, and the *thoroughness* as $\tau(\eta) = v(\eta)/v$. A test is efficient if it finds only a few correct (non-erroneous) values, and it is thorough if it finds most of the erroneous values. If the hypothesis is that a particular value is erroneous, then an efficient test is unlikely to lead to a Type I error (rejecting a correct value), for which the probability is $1 - \theta(\eta)$, but it may easily lead to a Type II error (accepting an erroneous value). The reverse is true of thorough tests.

The probability that a value chosen at random is erroneous is v/T, while the probability that one of the selected values is erroneous is $\theta(\eta)$. We define the *relative efficiency* of a test as the ratio of the two probabilities, or $\gamma(\eta) = \theta(\eta)T/v = Tv(\eta)/v\eta$. When, as is usual, v is unknown, θ can perhaps be estimated but τ and γ cannot. As η increases from zero, θ is initially zero or one, and $\theta \to v/T$ as $\eta \to T$. In order for a test to be better than random selection, $\gamma > 1$ is required for some range of η. If this occurs, it is not evident *a priori* where θ is maximal; and although θ eventually declines, the decline need not be monotonic. With increasing η, τ can be expected to rise monotonically. Balancing the effects on θ and τ, and the relative importance of Type I and Type II errors, leads in principle to an appropriate choice of η.

3. The Distribution of Errors

Let $x_1, \ldots, x_T$ be the true values of a variable, and let $x_1^{(r)}, \ldots, x_T^{(r)}$ be the corresponding reported values. Let α_t be a random variable with probability distribution $p(\alpha)$, independent of x_t. We assume that

$$x_t^{(r)} = f(\alpha_t)x_t \tag{1}$$

If $x_t^{(r)}$ is written as $x_t + v_t$, the additive error v_t is a random multiple of x_t. One complication is that if x_i is a component of x_j, and x_{it} contains an error, so does x_{jt}, with

$$x_{jt}^{(r)} = (1 + [f(\alpha_{it}) - 1]x_{it}/x_{jt})x_{jt} \tag{2}$$

so the error in x_{jt} is correlated with x_{it} as well as with x_{jt}.

Now let $s(x_1, \ldots, x_T)$ be any statistic to be calculated from the sample. Let ξ_ρ be the ρ-th moment of $f(\alpha)$, or

$$\xi_\rho = \sum_\alpha f^\rho(\alpha)p(\alpha) \tag{3}$$

where for simplicity we assume that α has a discrete distribution. If a statistic s is homogeneous of degree ρ in the values of x, then it follows that s calculated from $x^{(r)}$, and s calculated from x, are related in probability by ξ_ρ. That is, since

$$E[x_t^{(r)\rho}] = E[f^\rho(\alpha_t)x_t^\rho] = E[f^\rho(\alpha_t)]x_t^\rho = \xi_\rho x_t^\rho \tag{4}$$

it follows that

$$E[s(x_1^{(r)}, \ldots, x_T^{(r)})] = \xi_\rho s(x_1, \ldots, x_T). \tag{5}$$

The exact value of $s(x_1^{(r)}, \ldots, x_T^{(r)})$ depends of course on *which* values contain errors;

expression (5) merely emphasizes the usefulness of knowing something about the error distribution.

A particularly convenient form of $f(\alpha)$ is $f(\alpha_t) = \exp(\alpha_t)$. Then

$$\log x_t^{(r)} = \alpha_t + \log x_t \tag{6}$$

and

$$E[s(\log x_1^{(r)}, \ldots, \log x_T^{(r)})] = \mu_\rho + s(\log x_1, \ldots, \log x_T) \tag{7}$$

when the statistic s is hom-ρ in $\log x$. μ_ρ is the ρ-th moment of the distribution of α, or

$$\mu_\rho = \sum_\alpha \alpha^\rho p(\alpha) \tag{8}$$

If there are different types of errors (with α serving as an index of severity or frequency), a test may be efficient, or thorough, at finding some errors but not others. The number ν is replaced by the vector $\nu(\alpha)$, where $\nu(\alpha_t) = p(\alpha_t)T$ and $\sum_\alpha \nu(\alpha) = \nu$. Similarly $\nu(\alpha, \eta)$, $\theta(\alpha, \eta)$, $\tau(\alpha, \eta)$ and $\gamma(\alpha, \eta)$ are defined, where $\nu(\eta)$ and $\theta(\eta)$ can be found by summation over α but $\tau(\eta)$ and $\gamma(\eta)$ cannot. The design of a test should take account of which class(es) of error it is most important to detect; it may not matter if some errors go unnoticed. The functions $\theta(\alpha, \eta)$ need not move together with increasing η, for different values of α; nor need the $\tau(\alpha, \eta)$. This complicates the choice of η.

If (3) refers to all the errors initially in the data, we can define the moments corresponding to the errors remaining after applying a test of thoroughness $\tau(\alpha, \eta)$ as

$$\xi_\rho(\tau) = \sum_\alpha [1 - \tau(\alpha, \eta)] f^\rho(\alpha) p(\alpha) \tag{9}$$

If $f^\rho(\alpha)$ is replaced by α^ρ, expression (9) gives the moment $\mu_\rho(\tau)$. The importance of an error α depends on the function f, the frequency $p(\alpha)$ and the particular values x_t for which $\alpha_t = \alpha$.

4. Several Variables and Prior Information

Suppose that our prior information about a set of n variables can be expressed as

$$b_0 - x_{1t} + b_2 x_{2t} + \ldots + b_n x_{nt} + \varepsilon_t = 0 \tag{10}$$

where $b_1 = 1$ for normalization, and ε_t is an error term with zero mean and constant variance, independent of $x_{1t}, \ldots, x_{nt}$. Substituting the reported values, some of which contain errors,

$$b_0 - x_{1t}^{(r)} + b_2 x_{2t}^{(r)} + \ldots + b_b x_{nt}^{(r)} + \varepsilon_t + v_t = 0 \tag{11}$$

where v_t includes the effects of the errors in the variables, and may not be well-behaved. ε_t and v_t cannot be observed separately; only the sum $u_t = \varepsilon_t + v_t$ is observable. Since ε_t and v_t are uncorrelated, the larger is a value u_t, the more likely it is to contain a non-zero error v_t. The variance σ_ε^2 is unknown, so the best measure of "large" u_t is the variance $\sigma_u^2 = \sigma_\varepsilon^2 + \sigma_v^2$. Let k be a parameter describing the

stringency of the test: then an observation $[x_{1t}^{(r)}, \ldots, x_{mt}^{(r)}]$ is said to be *extreme* if $u_t^2 > k^2\sigma_u^2$.

5. Multivariate Tests

Suppose that $x_2, \ldots, x_n$ are observed without error (x_1 is the variable to be tested). Then

$$x_{1t}^{(r)} = b_0 + b_2 x_{2t} + \ldots + b_n x_{nt} + u_t \tag{12}$$

This relation can be estimated without bias by ordinary least-squares regression if $E(v_t) = 0$ and $E(v_t^2)$ is independent of x_{1t}: that is, if the errors v have the same characteristics as the errors ε. Otherwise $b_0, b_2, \ldots, b_n$ will be estimated with bias, as will the distribution of the errors u. The usual procedures for coping with heteroscedasticity, such as dividing all the variables by x_1 or x_1^2, are of no help since that would introduce errors on the right-hand side. It is also impossible to adjust for non-zero mean error if ξ_1 is unknown.

Since the object of the test is not to estimate $b_0, b_2, \ldots, b_n$, it may not appear to matter if they are biassed. It is important however that the regression provide a good fit to the sample. When the regression is not significant, the expected value for the dependent variable (x_1) is just the sample mean. Large values of u_t^2 are then associated with large (or small) x_t, that is, with values which are extreme *without* considering any other variables. In these circumstances the multivariate test collapses to a univariate test. Furthermore, biassed coefficients pull the regression line toward the erroneous values, which makes them harder to detect (by reducing their residuals) and makes some correct values appear erroneous. Therefore we consider three possible means of modifying a multivariate test so as to retain the relation (12) while reducing the bias likely to be introduced by OLS regression on the full sample.

The simplest procedure is to estimate (12) from a censored sample, excluding those observations most likely to contain large errors v_t. A large enough error in $x_{1t}^{(r)}$ will not only make u_t extreme, but will make $x_{1t}^{(r)}$ extreme compared to the other values of $x_1^{(r)}$, independently of the values of $x_{2t}, \ldots, x_{nt}$. The univariate extreme values should therefore perhaps be excluded. If the excluded values are in fact erroneous, this procedure will reduce σ_u^2, improve the estimates of $b_0, b_2, \ldots, b_n$ and (for a given value of k) make the test more stringent.

A second possibility is to use an estimating procedure which is relatively insensitive to large residuals, rather than least-squares estimation. The ideal regression method might estimate (12) by minimizing

$$\sum Q(x_{1t}^{(r)} - \hat{x}_{1t}), \quad \text{where} \quad \hat{x}_{1t} = \hat{b}_0 + \hat{b}_2 x_{2t} + \ldots + \hat{b}_n x_{nt} \tag{13}$$

and the function Q would have the properties $Q(0) = 0$, $Q(-x) = Q(x)$, $Q(x) \geq 0$, $Q'(x) \geq 0$ and $Q''(x) \leq 0$. Beyond some distance from the regression line, a point should cease to have any (further) influence on the estimates; so $Q''(x) \leq 0$ is desirable. Computing algorithms do not exist except for $Q(x) = |x|$. It is not vital to have a zero mean residual, since $E(v_t) \neq 0$ necessarily.

The third possibility is to retain all the sample points and use OLS estimation for the ease of computation, but to group the data before estimating. If the observa-

tions are appropriately grouped, the regression can be protected from individual errors.

A multivariate test may be justified to the extent that it (i) selects observations in the tail(s) of the distribution more efficiently than a univariate test, or (ii) finds erroneous values in the center of the distribution, which would escape a univariate test. The test need not be symmetric: a given η can be divided between too-high and too-low values of $x_{it}^{(\eta)}$ by testing $u_t > k_1\sigma_u$ and $u_t < -k_2\sigma_u$, for $k_1 \neq k_2$. The object in using a multivariate test is to trade assumptions about error distributions for assumptions about relations among variables, where the latter kind of information is more likely to be available. The relations to be tested can be based on, or even identical to, the relations to be examined after the data have been cleaned; using them at an earlier stage may tell something about their plausibility at the same time that errors are detected. Whether the additional cost of a multivariate procedure is repaid in greater efficiency or thoroughness is a question for empirical examination.

6. The Data Analyzed

In 1966–1972 household budget surveys were conducted in 18 major South American cities as part of the ECIEL Program coordinated by the Brookings Institution.[1] The data collected are in many cases the most complete or the most accurate available; nonetheless it was expected that they would include a variety of errors and would require careful cleaning before analysis.

The samples, and the procedures for treating the data, have been extensively described elsewhere [16], [17]. We indicate a few characteristics of several samples for which the cleaning process is (essentially) complete and from which some conclusions can be drawn. Some results were previously reported for the first sample studied [12]. All the stratified samples are non-proportional.

Country	*Cities*	*No. Observations*	*No. Intervals*	*No. Strata*
Colombia	4	2949	4	3
Chile	1	3378	4	3
Paraguay	1	568	2	(1)
Peru	1	1357	4	4

7. The Tests Applied

These data were subjected to two extreme-value tests. The first is a univariate frequency distribution which selects all the observations outside a specified range. The usual test was to define the range as $\bar{x} \pm 3\sigma_x$, with $\bar{x}$ (mean) and σ_x (standard deviation) estimated by first observing the entire distribution. The test has regularly been used only on the upper tail of the distribution: often $\bar{x} - 3\sigma_x < 0$, while $x > 0$ is required. The second test is a regression model of the form (12). The dependent variable was in most cases a share of total expenditure on a

[1] ECIEL is the Spanish acronym for Joint Studies on Latin American Economic Integration. 14 institutions in ten countries collaborated in this study; four are national statistical offices and ten are universities or private research institutes.

particular category of goods and services, and to minimize bias in the estimation due to errors in $x_2, \ldots, x_n$, all the latter were usually dummy variables. The usual criterion was $k = 3$, or $u_t > 3\sigma_u$: as with the univariate test, very few too-low values were detected.

8. Some Outcomes

Almost invariably there are a few very high values in the upper tail, with the highest observations exceeding $\bar{x} + 6\sigma_x$ and with approximately one or two percent of the observations exceeding $\bar{x} + 3\sigma_x$. The univariate test finds these extreme values quickly and cheaply, but it does not select any values in the center of the distribution. The very high values are quite frequently erroneous, or—when the same observation appears for several variables—come from an unrepresentative household. Once these values are eliminated or corrected, the test becomes much less efficient.

The regression test is considerably more expensive than the univariate procedure. The first questions of interest are (i) do the two tests select (essentially) the same observations, for equal η, and (ii) if they do not, is the multivariate test more efficient. The answers appear to depend very much on exactly how the tests are performed. If the share-of-expenditure is tested both ways, the two tests tend to pick out the same values. For example, in nine of the ten variables tested for Chile, the univariate test (at slightly higher η) found all the regression-test errors. Three multivariate tests for Peru yielded values always above $\bar{x} + 4\sigma_x$. Four such tests for Paraguay yielded six extreme values, of which four exceeded $\bar{x} + 7\sigma_x$ and two fell under $\bar{x} + 2\sigma_x$; four tests for Columbia detected only one value under $\bar{x} + 3\sigma_x$. Three other tests found 212 of 391 values below $\bar{x} + 3\sigma_x$. In all these cases, the residuals from the two tests are highly correlated.

If instead the regression test is based on share-of-expenditure while the univariate test is based on actual expenditure, the results are very different. Of 102 extreme values detected in 19 variables for Paraguay, only 45 had values above $\bar{x} + 3\sigma_x$, 42 were below $\bar{x} + 2\sigma_x$, and nine were below $\bar{x}$. 14 such tests for Peru yielded 152 extreme values, with 66 above $\bar{x} + 3\sigma_x$, 72 below $\bar{x} + 3\sigma_x$, and eight below $\bar{x}$. It is evident that the multivariate test can find extreme values which are hidden in the univariate distribution, and therefore that in general the analysis of any variable should take some account of the values of other variables. However, in the case of expenditure variables, a great deal is gained simply by taking ratios of total expenditure, after which the multivariate test adds relatively little. Also, it is not so valuable to select "extreme" values near the mean if most of the interior values are correct, and most of the errors lie several standard deviations away. In the Peruvian tests described, it was possible to correct 56 of the 152 values selected, and of these 39 exceeded $\bar{x} + 3\sigma_x$. The efficiency of the multivariate test averaged 0.33 overall, with 116 errors found in 354 values selected in 36 variables. In the Colombian sample the efficiency was 0.16, for 932 observations selected from 40 variables. In the Paraguayan sample almost no errors requiring correction were found. The rather poor performance of the multivariate test may be partly due to the use of dummy variables to explain a ratio with a fairly low variance in the sample. Probably more important is the fact that often

the regressions were not significant (by an F-test) so that the test collapsed to a univariate inspection.

These results are inconclusive, because differences between tests may be submerged by differences between types of variables tested or by an unsatisfactory specification of the regression. The true error distribution is unknown, so that τ cannot be estimated; for the same reason, the efficiency of both tests may be underestimated. An experiment was therefore conducted by deliberately introducing errors in the data from one sample, and then comparing the univariate and multivariate procedures for finding them. The errors are multiplicative, of the form $f(\alpha_t) = \exp(\alpha_t)$.

9. Design of the Experiment

Three distributions were used to generate errors in the Colombian sample. It was assumed that the artificial errors dominate, in number and severity, any errors remaining after the cleaning of the data. Distribution I was applied to three expenditure variables, Distribution II to six other expenditures, and Distribution III to two of those six.[2] The errors were carried into the logarithms of the variables and the shares of total expenditure. Observations were selected randomly with respect to city, interval and stratum. Either ten percent or four percent of the data were disturbed; this is believed greatly to exceed the true error frequency. The error probability distributions and their first and second moments are shown below:

α	-3	-2	-1	0	1	2	3
I $p(\alpha)$	0	0	0	0.9	0.1	0	0
II $p(\alpha)$	0.01	0.015	0.025	0.9	0.025	0.015	0.01
III $p(\alpha)$	0	0.0075	0.0125	0.96	0.0125	0.0075	0
	μ_1	μ_2		$\exp(\alpha)$		ξ_1	ξ
I	0.1	0.1				1.17	1.64
II	0	0.45				1.29	5.94
III	0	0.085				1.14	1.46

Both tests were then applied three times for each expenditure category: once to the actual value (EV), once to the logarithm (LEV) and once to the share in total expenditure (SEV). To improve the performance of the multivariate test, one continuous variable—total expenditure—was included among the independent variables. This exaggerates somewhat the efficiency of the test, since in practice such a variable might also contain errors.

10. Comparisons of Relative Efficiency

The statistic γ is used to compare the two tests. The results for the three variables affected by error-distribution I are as follows (for $k = 3.0$): results marked * are based on too few observations to be significant.

[2] The expenditures studied were: meat and fish, medical care, and household equipment and supplies (I); cereals, vegetables, clothing, personal care, education, and housing (II); and education and housing (III).

		univariate		multivariate	
		η	γ	η	γ
EV	1	55	8.1	65	8.6
	2	46	2.6	26	2.7
	3	44	2.7	33	3.6
SEV	1	24	8.8	50	8.0
	2	57	3.2	35	2.6
	3	27	6.7	50	6.6
LEV	1	27	1.5	36	0.6
	2	3*	3.3	2*	10.0
	3	51	1.0	5*	8.0

Tests using logarithms are almost useless for detecting asymmetric errors such as these. The tests of EV and SEV show, first, that γ varies considerably among variables; and second, that there is—at these values of η and v—no significant difference between the two tests. For such large multiplicative errors, an erroneous value is very likely to be extreme in the univariate distribution. We may suppose that with either increasing η or decreasing v, the multivariate test would improve its performance relative to the univariate test. Only further experiments, however, could show at what parameter values this would occur, and whether the gain would justify the additional cost.

A test based on Distribution I has the disadvantage that the results depend on the relation of $f(\alpha)$ to the range of x. Distribution II was introduced to minimize this problem and to see how well each test could pick out errors of one kind in the presence of errors of greater or lesser severity. If $x_t < x_{t'}$, $f(\alpha_t) > f(\alpha_{t'})$, and $x_t^{(r)} < x_{t'}^{(r)}$, a test should be more likely to select observation t than observation t'. A univariate test fails this criterion: the question then becomes whether a multivariate test can satisfy it.

The results of the comparison for the six affected variables are shown below, giving η and γ_α for $\alpha = -3, -2, -1, 1, 2, 3$. The other measure shown is the mean severity of the errors detected, defined as

$$\alpha^* = \log\left[\sum_\alpha \gamma_\alpha \exp|\alpha| / \sum_\alpha \gamma_\alpha\right] = 0 \tag{14}$$

This measure increases (but is ≤ 3.0) when errors are found at $\alpha = \pm 3$, and decreases as the errors detected are less severe (have lower values of α). It does not matter on which side of zero α lies. The statistic is of interest only when there are different kinds of errors in the distribution: α^* is uniformly 1.0 for error-distribution I.

The multivariate test appears to perform overall at least as well as the univariate test. Both tests concentrate on $\alpha > 0$ when the absolute expenditure or the share is analyzed; the univariate test is more likely to find errors with $\alpha < 0$. The regression test finds errors much more symmetrically when logarithms are examined. The regressions generally have R^2 between 0.2 and 0.5, with several coefficients significant: so the two tests really are different, although because of the large values of $f(\alpha)$ they find many observations in common. There does not seem to be any connection between the goodness of fit of a regression and whether it out-performed the univariate test. The differences in α^* and γ_α between the two

		γ_* Univariate Test								γ_* Multivariate Test							
		η	−3	−2	−1	1	2	3	α_*	η	−3	−2	−1	1	2	3	α_*
EV	1	39				1.0	18.9	65.6	2.84	78					20.1	63.9	2.84
	2	38				1.0	19.4	64.7	2.84	38					17.6	67.2	2.86
	3	10					20.1	59.0	2.83	8					16.8	73.7	2.88
	4	24				6.7	13.9	61.4	2.84	23				5.2	8.7	72.7	2.88
	5	15				2.7	13.4	65.6	2.86	10					6.7	78.6	2.95
	6	32			1.2	3.7	16.8	40.0	2.72	27				3.0	14.9	43.7	2.78
SEV	1	59			0.7	2.0	26.2	43.4	2.68	61				1.3	26.9	43.6	2.71
	2	54			0.8	0.8	22.3	49.2	2.76	61				1.3	26.3	43.6	2.72
	3	27				5.9	19.8	29.1	2.61	18				6.7	18.6	49.2	2.73
	4	53			1.5	4.5	21.5	50.0	2.72	52				1.5	25.8	49.2	2.74
	5	52				6.9	12.9	35.9	2.71	35				3.4	19.2	50.5	2.76
	6	44			0.9	3.6	13.7	44.7	2.78	47				2.6	12.8	39.7	2.79
LEV	1	71	30.4	6.6	1.1	1.1	8.5	34.6	2.85	90	26.2	8.9		0.4	13.4	29.5	2.80
	2	63	25.0	4.2	1.3	1.3	9.6	35.9	2.84	86	21.7	7.8	0.4		14.8	30.9	2.78
	3	0								13	7.6			3.1	5.2	60.5	2.92
	4	13				3.1	15.5	68.1	2.84	55	10.7	4.9		0.7	14.6	44.7	2.82
	5	4					16.8	73.8	2.88	19	15.5	3.6			7.0	56.9	2.92
	6	6						81.9	3.00	50	25.6	2.7		2.4	2.7	29.5	2.91

procedures are so small that it is not clear the greater cost and complexity of the multivariate test are justified.

In three respects, this comparison is unfair to the multivariate test. First, the experiment was limited, particularly by having $\eta \ll \nu$ in all cases. Second, the regressions are ordinary least-squares, and therefore suffer from the biasses described in section 5 above; also, all the observations were used, without grouping. Third, both tests were applied to the identical data, which included some extreme values easily detected by the univariate test. To the extent that the multivariate test "wasted its time" in finding those errors, it was less able—for a given value of k, or of η—to detect errors buried in the center of the distribution. The regression method would probably be much more efficient, relative to the univariate inspection, if the univariate extreme values were first removed.

Error-distribution III was introduced to reduce the total number of errors and their maximum severity, so as to reduce the importance of the first and third problems just described. Errors of $\alpha = \pm 3$ were eliminated, and $p(\alpha)$ was halved for $\alpha = -2, -1, 1, 2$. Four percent errors remained in the data. The stringency was also varied, to see the effect of changing η: values of k of 2.5 and 3.5 were used. This distribution was applied to variables 5 and 6 only: the results are shown below.

		γ_α Univariate Test						γ_α Multivariate Test				
	η	-2	-1	1	2	α^*	η	-2	-1	1	2	α^*
						$k = 2.5$						
EV 5	78			5.1	8.6	1.73	57			7.0	16.5	1.79
6	38			6.3	17.6	1.82	22			10.8	30.4	1.82
SEV 5	67			5.9	12.0	1.76	37			10.8	21.8	1.77
6	50			8.0	18.8	1.79	48			5.0	28.0	1.90
LEV 5	36	7.5				2.00	30	9.0		5.3	22.3	1.90
6	46		3.5		5.8	1.73	71	15.1			18.9	2.00
						$k = 3.5$						
EV 5	48			6.6		1.00	27			8.8	9.9	1.64
6	19			8.4	28.3	1.85	12			6.6	44.6	1.91
SEV 5	33			9.6	16.2	1.73	23			17.3	29.1	1.73
6	17			4.7	23.6	1.89	19				42.4	2.00
LEV 5	2						1					
6	0						21	6.4			19.2	2.00

Under these circumstances, the multivariate test performs better relative to the univariate test. For $k = 2.5$, it yields lower η and higher γ and α^* in almost every case. The increased efficiency is not at the expense of thoroughness; η can be lower while still detecting a large share of the errors in the data. At $k = 3.5$, η is about halved for the expenditure and share variables, but drops almost to zero for most of the logarithms. In general, the superiority of the multivariate test is more pronounced than at the lower stringency. It is much more efficient than the univariate test at finding the large errors ($\alpha = 2$). It appears that the regression procedure is superior when there are errors of different degrees of severity in the data; when the most severe errors present are still not so large as always to lead to univariate extreme values; and when the total number of errors is not too large. In these conditions, the multivariate test is markedly more efficient at detecting the more severe errors, and—when examining a small number of observations—

more efficient overall. When these conditions do not hold, some prior examination of the univariate extreme values appears to be desirable. Improved versions or ways of using a multivariate test should increase these advantages by reducing the estimation biases and allowing the test to hunt for errors in the center of the distribution of the variable examined.

11. Bibliographic Note

The information collected in household budget surveys may be thought of as generated by a sequence of steps, each of which allows the introduction of errors. Initially there are response errors, due to incomprehension, deceit or forgetfulness. Subsequently the data may be incorrectly coded or keypunched. Errors can also arise if values must be converted to different physical or monetary units or periods of reference. Some true values, containing none of these errors, may also be so unrepresentative that they might better be considered erroneous. All these difficulties increase when several slightly different samples are to be compared, so that more stages are required to harmonize them.

In principle, most of the errors created *after* a household is interviewed can be prevented by sufficient care in designing questionnaires, training interviewers and verifying the field work and subsequent data manipulation. In practice, such care is not always taken. There are then two broadly-defined possibilities for analyzing the data (excluding the course of taking no account of the errors):

(1) Estimation of particular relations by models which expressly characterize the errors but do not identify them or remove them from the data; or

(2) Selection of certain values which are thought likely to be erroneous and which are then eliminated or replaced by information derived from the sample.

The latter procedure also requires that some assumptions be made about the errors so as to identify values which are likely to be in error. We assume that systematic errors can be corrected at an early stage in the analysis, so that the remaining errors affect a small share of the observations. Errors in qualitative variables can often be detected with the aid of strong prior information. Only certain (coded) values of a variable may be allowed, or only certain logical relations with other variables. For quantitative variables, however, the only prior restriction may be nonnegativity, and an error may lead to an extreme value which will bias any calculation based on that variable.

Procedure (1) is the domain of the errors-in-variables model (EVM) [13, chapter 10]. Provided one can estimate the covariance matrix of their errors, any combination of variables can be used for linear regression analysis. Since the model assumes zero means for all the errors, analyses based on mean values, such as tabulations, are unbiased. The errors are also assumed to have constant variances and to be independent of the true values of the variables. These assumptions may apply to conceptual variables such as permanent income [9], but they do not plausibly characterize the errors obtained in survey data. Such errors do not appear generally to have zero means [7], [8] and even when they are symmetric and have small means they may be correlated with the variable in which they occur or with related variables [1]. The assumptions seem not to hold exactly even for data much less subject to error than those in household surveys [15].

Because both dependent and independent variables contain errors, it is not possible in this model to estimate individual errors without the additional restrictions that the covariance matrix be diagonal and that each true variable be an exact function of some exogenous, error-free variables [10]. Even if all the restrictions can be accepted, any nonlinear transformation of the data will change the error structure. If the object is not only to estimate certain relations but to leave the data ready for other analyses, this procedure is not of much help.

The problem becomes much simpler if only the dependent variable is assumed to contain errors. Then it may be possible to estimate individual errors; and even if this is not done, both the true relation and some parameter(s) of the error distribution may be estimated without bias. The assumptions of independence, zero mean and constant variance may be dropped. An example is Elashoff's model [6], in which the dependent variable includes errors which are quadratic functions of the independent variable. (The regression line could be used to impute true values, if desired.) Chen and Dixon [2] consider the dependent variable to include a normal error in either location or scale. For a certain range of probabilities of error, it is shown that either trimming or Winsorizing the set of values of the dependent variable associated with *each* value of the independent variable, gives better estimates of the regression coefficients than are obtained by ignoring the errors. The improvement disappears as the probability of error rises.

Such adjustments are already an example of method (2). Many procedures proposed for data editing are of this form: certain values are either eliminated or changed, without verifying the existence or size of an error. Often they are designed to improve the estimation of some statistic(s) by eliminating or reducing the influence of the errors. An example is McCarthy's [14] suggestion for discarding "inliars" to improve the dichotomous classification of a variable; another is Searls' [18] proposal to reduce the effect of large true values on the estimate of the mean. A number of contributions such as [5] discuss parameter estimation for particular distributions—most often the normal—when some values are erroneous or missing. The distribution of the errors is still often assumed to be normal. A general procedure for dealing with outliers or with a long-tailed error distribution is presented by Tukey [19, pp. 21–32].

Further analysis, and the identification of individual errors, often is feasible if (i) the data have passed through several stages, and it is possible to check a doubtful value against an initial entry, or (ii) the verification of errors can draw on information in the sample or exogenous to it, which was not used to select the observations for analysis. Both conditions are likely to hold for consumer survey data; (i), because data that have been coded, converted and keypunched can be compared to questionnaire entries, for correction of errors introduced at these stages, and (ii), because the number of variables is likely to be much too large to use them all in the selection procedure.

Much of the literature on the detection of errors (for example [3], [4], [11]) is characterized by the following set of assumptions:

1. Only one variable is considered.

2. Errors in the variable are most likely to give rise to outliers, so the test should determine whether the highest (or lowest) values are erroneous.

3. The errors are normally distributed and independent of the true values of the variable.

4. The sample is small (often ≤ 20) and only one or a few outliers are to be tested.

5. The true distribution of the variable is known (often normal) and the chief problem may be to estimate its parameters in the presence of errors.

Clearly not all these assumptions apply to all the procedures available, but some subset of them nearly always appears.

In this inquiry, we abandon assumptions 4 and 5. Assumption 1 is (largely) retained. Assumptions 2 and 3 are special cases of more general hypotheses put forward about the errors being sought. The model developed is somewhat similar to the balancing of costs-of-inspection and losses-from-errors discussed by van der Waerden for problems of quality control [20].

The Brookings Institution

References

[1] Michael E. Borus, "Response Error in Survey Reports of Earnings Information," *Journal of the American Statistical Association 61*, September 1966, 729–738.

[2] Edwin H. Chen and W. J. Dixon, "Estimates of Parameters of a Censored Regression Sample," *Journal of the American Statistical Association 67*, September 1972, 664–671.

[3] H. A. David and A. S. Paulson, "The Performance of Several Tests for Outliers," *Biometrika 52*, 1965, 429–436.

[4] W. J. Dixon, "Processing Data for Outliers," *Biometrics 9*, 74–88.

[5] W. J. Dixon, "Simplified Estimation from Censored Normal Samples," *Annals of Mathematical Statistics 31*, June 1960, 385–391.

[6] Janet D. Elashoff, "A Model for Quadratic Outliers in Linear Regression," *Journal of the American Statistical Association 67*, June 1972, 85–87.

[7] Robert Ferber, "The Reliability of Consumer Surveys of Financial Holdings: Time Deposits," *Journal of the American Statistical Association 60*, March 1965, 148–163.

[8] Robert Ferber, "The Reliability of Consumer Surveys of Financial Holdings: Demand Deposits," *Journal of the American Statistical Association 61*, March 1966, 91–103.

[9] Milton Friedman, *A Theory of the Consumption Function*, Princeton: Princeton University Press—National Bureau of Economic Research, 1957.

[10] Arthur S. Goldberger, "Maximum-Likelihood Estimation of Regressions Containing Unobservable Independent Variables," *International Economic Review 13*, February 1972, 1–15.

[11] Frank E. Grubbs, "Procedures for Detecting Outlying Observations in Samples," *Technometrics 11*, February 1969, 1–21.

[12] Howard Howe and Roberto Villaveces, "Data Preparation for Latin American Comparisons of Consumption," Conference on the Role of the Computer in Economic and Social Research in Latin America, Cuernavaca, Mexico, 25–29 October 1971 (mimeo).

[13] Edmond Malinvaud, *Statistical Methods of Econometrics*, Chicago: Rand McNally, 1966.

[14] Philip J. McCarthy, "The Effects of Discarding Inliars When Binomial Data are Subject to Classification Error," *Journal of the American Statistical Association 67*, September 1972, 515–530.

[15] Tracy W. Murray, "An Empirical Examination of the Classical Assumptions Concerning Errors in Data," *Journal of the American Statistical Association 67*, September 1972, 530–537.

[16] Philip Musgrove, "The Collection and Interpretation of Household Income and Expenditure Information," The Brookings Institution, May 1972 (mimeo).

[17] Philip Musgrove and Howard Howe, "ECIEL, Estudio de Consumo e Ingreso Familar: Antecedentes y Metodología," The Brookings Institution, April 1973 (mimeo).

[18] Donald T. Searls, "An Estimation for a Population Mean which reduces the effect of Large True Observations," *Journal of the American Statistical Association 61*, December 1966, 1200–1204.

[19] John W. Tukey, "The Future of Data Analysis," *Annals of Mathematical Statistics 33*, January 1962, 1–67.

[20] B. L. van der Waerden, "Sampling Inspection as a Minimum Loss Problem," *Annals of Mathematical Statistics 31*, June 1960, 369–384.