THE ECONOMIC
RESEARCH INSTITUTE
MEMORANDUM SERIES
NO. 72

Identification of Individual Abnormalities in Least Squares Regression

by R. C. Geary

Confidential: Not to be quoted until the permission of the Author and the Institute is obtained.

Identification of Individual Abnormalities in Least Squares Regression

by R. C. Geary

In single equation LS regression the common practice is to test goodness-of-fit by the standard error of estimate s and probable absence of residual auto-regression by the Durbin-Watson d [2], [3], or the more recent count of sign changes τ [6]. With a wide choice of causitive (or independent) variables (indvars) and with access to a computer, several regressions can be produced, one for each set of indvars selected. We usually pick the regression with the lowest s and a satisfactory d or τ as the "best", unless there are very compelling a priori reasons for picking some other set. Truth to say, there is still much empiricism in regression practice; in it art has a place as well as science.

In setting up the model $y = x\beta + u$, u regular, we are saying that (considering time series for convenience) throughout the period certain causes (which need not be independent) regularly affected the level of the dependent variable, the difference between the calculated vector $x\beta$ and the observed y vector, namely u, summarizing a vast number of unidentified causes, operating perhaps in some "years" but not in others, as well as plain errors of observation. We customarily regard our table of data as a single realization from a theoretically possible infinity of states (with x constant throughout), the minor causes operating in such a random way that the elements of u can be assumed to be homoskedastic throughout and independent of one another, i.e. u is regular, by definition.

It is customary (indeed a practice to be recommended) to graph the calculated dependent variable x b against the original vector y. Inevitably some of the calculated disturbances, elements of v, $y = x\,b + v$, are comparatively large. Are they abnormally so? The statistics s, d and r tell us little or nothing about such abnormality. Clearly we have something to gain by studying the individual disturbances. Our knowledge of the data will be deepened thereby. Such exercises may even bring to light causative variables then-to-fore unsuspected. When, and only when, abnormality has been stochastically determined are we justified in using the device of dummy variables thereby mitigating the effect of abnormality. This paper deals with the problem of the identification of such abnormalities.

Order Statistics

We do so by recourse to order statistics. The elements of u in the model are, by definition, independent. The vector v is an unbiased estimate of u. It may be assumed that number of sets of observations T is so large that the calculated elements of v are also independent; they cannot be so, in general, since $v' i_T = 0$, i_T the unit vector.

We deal throughout with absolute values of the disturbances and that each of these (positive) values has the cumulative frequency (c.f.) F, $0 \leq F \leq 1$. The c.f. of the absolute value of order n, on the null-hypothesis, is :

$$(1) \quad G_n = T \binom{T-1}{n-1} \int_0^F d\,x\,(1-x)^{n-1}\,x^{T-n}$$

as is well-known. For any order n the value of G_n as a

polynomial in F can easily be found by expanding $(1 - x)^{n-1}$ in the integral. The formulae for the first three orders are :

$$G_1 = F^T$$

$$(2) \quad G_2 = T F^{T-1} - (T - 1) F^T$$

$$G_3 = \frac{T(T-1)(T-2)}{2} \left[\frac{F^{T-2}}{T-2} - \frac{2 F^{T-1}}{T-1} + \frac{F^T}{T} \right]$$

The practical problem is : given G_n to find F. The solution is obvious in the case of G_1. In general G_n will have values like .95, .99 etc depending on the probability used for determining significance. Now for these earlier orders it is evident that, if $G_n = 1 - g_n$ and $F = 1 - f$, $f \ll g_n$, in fact f is very small. On making these substitutions we find from (2) :

$$(3) \quad \begin{aligned} &(i) \quad g_1 = \binom{T}{1} f - \binom{T}{2} f^2 + \binom{T}{3} f^3 - \ldots \\ &(ii) \quad g_2 = \binom{T}{2} f^2 - 2 \binom{T}{3} f^3 + 3 \binom{T}{4} f^4 - \ldots \\ &(iii) \quad g_3 = \binom{T}{3} f^3 - 3 \binom{T}{4} f^4 + 6 \binom{T}{5} f^5 - \ldots \end{aligned}$$

In general :

$$(4) \quad g_n = \sum_{i=0}^{T-n} (-)^i \binom{n+i-1}{n-1} \binom{T}{n+i} f^{n+i}.$$

Now, taking a line from the obvious advantage of such a course in the case of $n = 1$ we make such a transformation as to ensure that the first coefficient on the right of (4) is always unity. For this we set :

$$(5) \quad \binom{T}{n} f^n = x^n$$

or :

$$(6) \quad f = \binom{T}{n}^{-\frac{1}{n}} x.$$

On so substituting for f in (4), the typical term contains the following factor in T :

$$V_i^n = \binom{T}{n+i}\binom{T}{n}^{-\frac{(n+i)}{n}}$$

$$(7) \qquad = \frac{T\,\overline{T-1}\,\ldots\,\overline{T-n-i+1}}{(n+i)!}$$

$$\left[\frac{T\,\overline{T-1}\,\ldots\,\overline{T-n+1}}{n!}\right]^{-\frac{(n+i)}{n}}$$

easily seen to be $O(T^o)$ in T so that as $T \to \infty$ each transformed coefficient (i.e. of x^{n+i}) tends to a constant value, which on combining factors is easily seen to be :

$$(8) \qquad C_i^n = (-)^i \binom{n+i-1}{n-1} \left[(n+i)!\right]^{-1} (n!)^{\frac{n+i}{n}}$$

As the factors in T alone tend to unity when $T \to \infty$ C_i^n is the coefficient of x^{n+i} in transformed (4). For all values of n, $C_o^n = 1$, from (8). As an example, for the third term of g_3 at (3), i = 2, n = 3. Then, from (8) :

$$C_2^3 = (-1)^2 \binom{4}{2} (5!)^{-1} (3!)^{5/3}$$

$$= 6 \times (120)^{-1} \times (1.817121)^5$$

$$= 0.990579.$$

This is the value of the coefficient of x^5 in g_3 when number of sets of observations is indefinitely large.

In (7) the factors independent of T have been transferred to C_i^n, given by (8). The rest of (7) (i.e. the part in T alone) may be written, with $w = 1/T$:

$$(9) \qquad W_i^n = (1-w)(1-2w)\ldots(1-mw)$$

$$[(1-w)(1-2w)\ldots(1-rw)]^{-p},$$

$$m = n+i-1; \quad r = n-1; \quad p = (n+i)/n$$

The full coefficient of x^{n+i} in the x-transformed version of (4) will then be $C_i^n W_i^n$.

$\text{Log}_e\ W_i^n$ is expanded in powers of w. No special interest attaches here to the general expansion so we proceed to particular cases. Again we take i = 2, n = 3, as an illustration. Recalling that, when $0 \leqslant w \leqslant 1$, $\log_e (1 - w) = -w - w^2/2 - w^3/3 - \ldots$ ad inf we find :

$$\log_e W_2^3 = -5w - \frac{65}{6} w^2 - \frac{85}{3} w^3 - \frac{977}{12} w^4 - 249 w^5 -$$

When T = 100, $w = 10^{-2}$ and the first four terms of the expansion give $\log_e W_2^3 = -.051116$. Whence $W_2^3 = .95016$. The full coefficient of x^5 in the expansion of g_3 is $C_2^3 W_2^3 = 0.990579 \times 0.95016 = 0.9412$.

The object of the x-transformation is to ensure that the coefficients of x^{n+i} remain small as T increases, in the transformed version of (4). Given T, they also diminish as i increases. The fact that in the expansion of $\log_e W_i^n$ the coefficient of w^k tend to increase sharply with k is really no embarrassment, since w^k becomes very small as k increases when $T \geqslant 10$, so that only the first few terms are required for a close approximation. The x-transformed equations were used throughout for the computation of Table 1.

THE HIGHEST DEVIATE

On transformation (6) the x-equation version of (3)(i) is :

$$(10)\quad g_1 = x - \frac{T(T-1)}{1.2\ T^2} x^2 + \frac{T(T-1)(T-2)}{1.2.3\ T^3} x^3 - \ldots \text{ T terms.}$$

When $T = \infty$, (10) becomes :

$$(11)\quad g_1 = 1 - e^{-x}$$

yielding the solution :

$$(12) \quad x = g_1 + g_1^2/2 + g_1^3/3 + \ldots \text{ad inf.}$$

When $g_1 = .05$, $x = .051292$, when $g_1 = .01$, $x = .010050$. At the other extreme when $T = 1$, $x = g_1$. It is obvious that when g_1 is small the universal solution (i.e. for all values of T) is approximately $x = g_1$. If x_T be the exact solution the corresponding f-probability is found from (8) :

$$(13) \qquad f_T = x_T/T.$$

The normal theory g_1-probability null hypothesis critical point is that found from the standard normal table [1] corresponding to probability $(1 - f_T/2)$.

The Critical Probability Points Table

While the derivation of probabilities f corresponding to any initial probability g can easily be derived by the foregoing x-transformation, the derivation of critical points corresponding to probability f presents certain difficulties when T is not large. We use normal theory throughout this paper but such a procedure is not strictly valid. In the first place, even if model residue u_t is normally distributed, the statistic we use, namely v_t divided by its estimated standard error, is not distributed normally, but as the Student-Fisher statistic t. The hypothesis of normality is strictly true only as $T \rightarrow \infty$. In practice, however, Student-Fisher critical points, given probability (.05, .01 etc) are close to normal theory points and the values given in Table 1 can be used with the mental reservation that the actual null hypothesis probabilities are slightly greater than the .05 and .01 indicated. This is really an unimportant

point since we make only formal use of these probabilities in making inferences: we are content to state merely that some calculated value is "significant". It is enough that the null hypothesis probability is "small".

Another difficulty is that while the sample of T may be random and drawn from a normal population the statistics of given order are not normally distributed. The critical points of the statistic $X_1 = (x_1 - \mu)/\sigma$, where μ and σ pertain to the normal population sampled and x_1 the highest value in the sample, are given in Table 24 of [1], for $T \leq 30$, presumably using the exact frequency distribution of X_1. It remarkably happens, however, that our .05 and .01 probability critical values for sample sizes T = 10, 20, 30 (six values in all), though computed on the assumption that the largest value was normally distributed, exactly (to two decimal places) equal the [1] values. So much for smallish values of T. As T increases, the frequency distributions of statistics of all orders tend towards normality so that, for all values of T shown in Table 1, considerably confidence may be reposed. They are, however, described as "approximate" because the populations involved are really not normally distributed, but only approximately so, as explained above.

Table 1 is fundamental for the present research. As already stated, the values shown were derived using the x-equations. It was usually possible to make a good guess of a near approximation x_0 to the root x required, beginning the iterative process. Then $x_1 = x_0 + e_0$ where $e_0 = -f(x_0)/f'(x_0)$, $x_2 = x_1 + e_1$ etc. In fact, two iterations were required in only a few cases; mostly one sufficed since, as T increased, the x-solutions became

Table 1.

Approximate Critical .05 and .01 Probability Points of Deviates of Orders 1, 2, 3 (Absolute Values), for Certain Values of Sample Size T.

Normal Theory Assumed Throughout.

T	Order 1		Order 2		Order 3	
	.05	.01	.05	.01	.05	.01
1	1.96	2.58	-	-	-	-
10	2.80	3.29	2.09	2.42	1.71	1.98
20	3.02	3.48	2.36	2.67	2.03	2.28
40	3.22	3.66	2.61	2.90	2.31	2.54
60	3.33	3.76	2.75	3.02	2.46	2.69
80	3.41	3.84	2.84	3.11	2.56	2.78
100	3.48	3.89	2.91	3.18	2.64	2.85

nearly identical for each of the three orders. Critical points for only a few values of T are given since in the range $10 \leq T \leq 100$ critical points for any T can easily be interpolated.

Application to Regression

The λ_t factors. If the model is $y = x\beta + u$, u regular, the LS estimate is $y = y_c + v$, $y_c = x b$. Now, while $E u_t^2 = \sigma^2$ for all elements of u, the value of $E v_t^2$ is not σ^2 but $\lambda_t^2 \sigma^2$, where :

$$(14) \qquad \lambda_t^2 = 1 - x_t' (x' x)^{-1} x_t .$$

The second term on the right of (14) is $O(T^{-1})$ so that $\lambda_t^2 \to 1$ and $E v_t^2 \to \sigma^2$ only as $T \to \infty$. This λ-correction should be used when T is not large, as in the exercise that follows. In simple regression (14) assumes the form :

$$(15) \qquad \lambda_t^2 = (T - 1)/T - (x_t - \bar{x})^2 / \sum_{t=1}^{T} (x_t - \bar{x})^2 .$$

The null hypothesis here is that all the elements in the residual vector u have the regular properties $E\,u_t = 0$, $E\,u_t^2 = \sigma^2$ (with $E\,u_t\,u_{t'} = 0$, $t' \neq t$). We establish the regression on these assumptions about our model (or population), estimating σ^2 by $s^2 = \Sigma(y - y_c)^2/(T - K)$, where K is the number of indvars, including the constant. But if some of the values of the observed y_t contain abnormalities which we hope to discern from an examination of the higher values of $|v_t|$, we contemplate the possibility of a model in which a few of the disturbances are not, say, u_1, u_2, ... but $u_1 + \alpha_1$, $u_2 + \alpha_2$, ... so that all the disturbances are no longer regular, as defined above. Here we assume that these disturbances are few, say two or three: general residual heterovariance is another matter (see, for example, [4]).

If we use the larger $|v_t|$ to this end, we cannot safely use the classical formula in the previous paragraph, pace the null hypothesis, to estimate s^2 unless T is very large, which it rarely is. As the following exercise clearly shows, such an estimate, when T is not large, can seriously overestimate σ^2 and underestimate the test statistic, namely $v_t/(\text{s.e. } v_t)$, used for identifying abnormalities, thereby concealing these when they are possibly present. Nor, if, say, the first and second highest absolute values of the residuals are under test, simply to eliminate from sum squares (and reducing d.f. by 2) is invalid, for, even if there were no abnormalities present, this procedure would underestimate σ^2.

The correct treatment is given in [5] . This involves substituting statistics $s^2\,E_1$, $s^2\,E_2$ etc for the actual contributions of the highest, second highest etc $(y_t - y_{tc})^2$ in sum squares, where $E_n = E\,z_n^2$, z_n being the

ith order value for a random sample of T from $N_+(0, 1)$. We then estimate σ^2 by s^2 from :

$$(16) \quad (T - K)\, s^2 = s^2 \sum_{n=1}^{\kappa} B_n + \Sigma' (y_t - y_{tc})^2 ,$$

where Σ' indicates the omission of the κ residuals under test, or :

$$(17) \quad s^2 = \Sigma' (y_t - y_{tc})^2 / (T - K - \Sigma E_i).$$

To make the present <u>exposé</u> tolerably complete, a table of E_i, given in [5] is reproduced here, for convenience, as Table 2.

TABLE 2

Value of $E_n = Ez_n^2$ for random samples of T from $N_+(0, 1)$ for $\sigma^2 = 1$

Sample size T	*Descending order* n							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
10	3·799621	2·171462	1·426472	0·970990	0·660253	0·437538	0·275135	0·155713
20	4·916871	3·216540	2·410593	1·897855	1·528207	1·245702	1·020668	0·836765
30	5·599340	3·867966	3·037613	2·502189	2·112625	1·809929	1·564854	1·360810
40	6·093230	4·343362	3·498975	2·951316	2·550458	2·237010	1·981502	1·767200
50	6·480929	4·718344	3·864523	3·308782	2·900577	2·580232	2·318119	2·097405
60	6·800321	5·028251	4·167506	3·605907	3·192432	2·867188	2·600425	2·375213
70	7·072022	5·292497	4·426376	3·860271	3·442774	3·113818	2·843555	2·615017
80	7·308510	5·522905	4·652444	4·082727	3·662028	3·330132	3·057110	2·825948
90	7·517919	5·727220	4·853153	4·280453	3·857122	3·522820	3·247552	3·014259
100	7·705850	5·910793	5·033661	4·458440	4·032894	3·696576	3·419431	3·184363

<u>Effect of abnormalities on the regression</u>. Since $b - \beta = (x' x)^{-1} x' u$, if a finite number of abnormalities are present in u, b is no longer an unbiased estimate of β. It is, however, a consistent estimate. For example, in simple regression $b - \beta = \Sigma(x_t - \bar{x})\, u_t / \Sigma (x_t - \bar{x})^2$. The point is that in the formula for b the biases, if any, enter linearly (instead of their squares in estimating σ^2 by the classical formula) and in practice the bias in estimate of β will usually be small, so that this biased

estimate will lie well within the confidence limits of impeccable estimate.

An Exercise

The regression is simple :

$$Y_t = 10 + x_t + u_t,$$

$x_t = -13, -12, \ldots 0, \ldots, 12, 13$, $\beta = 1$, $T = 27$ and the u_t initially a random sample of 27 from N (0, 1). Disturbances of 4 were added to the y_t corresponding to $x_t = -9$ and to that for $x_t = 4$. The (x, Y) "observations" are shown as dots on the chart. The problem: can we detect these two abnormalities (clear enough to the eye) by stochastic analysis, and correct for them? For straightforward LS we have : $T = 27$; $\Sigma x = 0$; $\Sigma x^2 = 1,638$; $\Sigma x\, Y = \Sigma x\, y = 1,678.45$; $\Sigma Y = 278.45$; $\Sigma Y^2 = 4,666.9647$; $\Sigma y^2 = 1,795.3210$. From these we find a (estimate of the interceptof the regression) = 10.3130 and b = 1.0247, obviously very good estimates of the population values 10 and 1 and illustrating the point in the foregoing text that good regression coefficient estimates can be obtained from disturbed data.

It is quite otherwise with the original estimate of s^2 (25 d.f.) which turns out to be 3.0169, three times the population value!

The two largest disturbance values v_t (as we might expect) are :

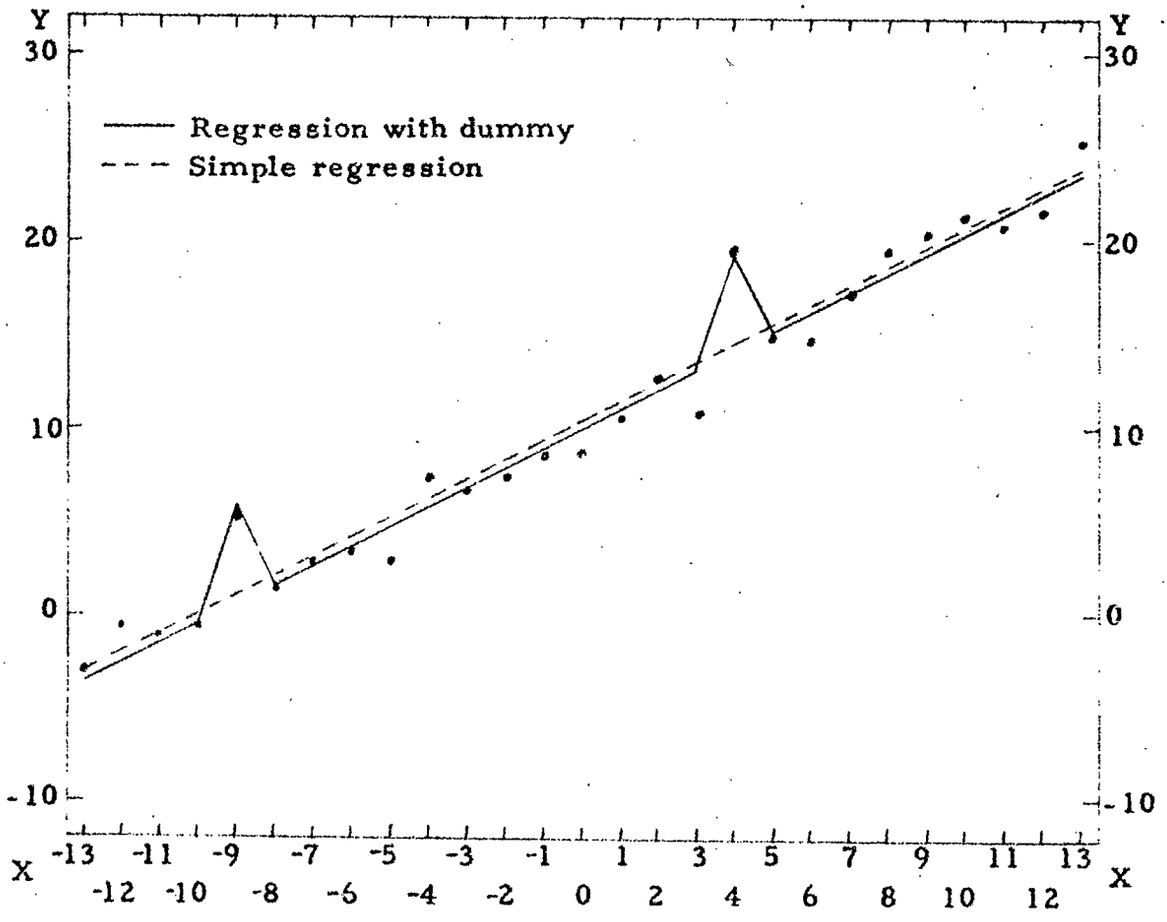
5.35 for x = 4
4.14 for x = -9

We try to show that these are abnormal by the method indicated in the text proper. We find :

$$\Sigma' = 75.4217 - (5.35)^2 - (4.14)^2 = 29.6596.$$

CHART

Constructed illustration. Data, (i) simple regression and (ii) regression with dummy variable.



By rough interpolation from Table 2 we find for T = 27, $B_1 = 5.4$, $B_2 = 3.7$, sum 9.1, so, from (17) :

$$s^2 = 29.6596/(25 - 9.1) = 1.8654$$

which, by chi-squared, is not significantly different from unity. $s = 1.3658$.

We then have :

x	Y	Y_c	v	λ	$v/\lambda s$
4	19.76	14.41	5.35	.9763	4.01
-9	5.23	1.09	4.14	.9558	3.17

From Table 1 the last column entries are well above the .01 probability points for orders 1 and 2 respectively. Abnormalities, we infer, are probably present.

The third highest value of $\left|v/\lambda s\right|$ is 2.03 which is clearly lower than the .05 probability point for order 3, T = 27, so we infer non-significance. We have justified the correction or elimination of the two variables.

Let the dummy variable be X_1, taking the values 1 for x = -9 and x = 4, otherwise 0. For the regression of Y on x and X_1 we require, in addition to the values already given :

$$\Sigma X_1 = 2; \quad \Sigma X_1^2 = 2; \quad \Sigma x_1^2 = 1.851852; \quad \Sigma x_1 y = 4.364080.$$

The regression is :

$$Y_c = \underset{(47.2)}{9.9303} + \underset{(40.0)}{1.0405}\, x + \underset{(6.7)}{5.1659}\, X_1 ,$$

the figures in brackets being the Student-Fisher t's. The population values of the first two coefficients, known to be 10 and 1, lie comfortably within the .95 probability confidence limits of estimate. Y_c is graphed as the firm

line on the chart. We find $s^2 = 1.1004$, also near the population value of unity. This latter result is flattering to the theory, since we deliberately took the two abnormal disturbances as equal. In practice this may not be the case and, if more than one abnormality is present, the estimate s^2 will be inflated. However, the s^2 will be liable to be much smaller than it would have been if the original data were regressed uncorrected; and a major objective of regression is the reduction of s^2. If more than one abnormality is detected, it will be sensible in the dummy to use +1 for positive and -1 for negative abnormality, otherwise 0.

Obviously the dummy variable procedure for correction is imperfect and other methods are conceivable, including complete elimination, though such a course is distasteful. We leave this aspect to others. The main object here is detection of abnormality, rather than its elimination.

****

31 March 1971.

REFERENCES

[1] Biometrika Tables for Statisticians (1958) Ed. E.S. Pearson & H.O. Hartley. Second Edition. Cambridge University Press.

[2] Durbin, J. & Watson, G.S. (1950). Testing for serial correlation in least squares regression. I. Biometrika, 37, 409-27.

[3] Durbin, J. & Watson, G.S. (1951). Testing for serial correlation in least squares regression. II. Biometrika, 38, 159-78.

[4] Geary, R.C. (1966). A note on residual heterovariance and estimation efficiency in regression. The American Statistician, October 1966, 30-31.

[5] Geary, R.C. (1967). Ex post determination of significance in multivariate regression when the independent variables are orthogonal. The Journal of the Royal Statistical Society, Vol. 29, No. 1, 154-61.

[6] Geary, R.C. (1970). Relative efficiency of count of sign changes for assessing residual autoregression in least squares regression. Biometrika, 57, 1, 123-27.