Residual Heterovariance and Estimation Efficiency in Regression

By

R. C. Geary

Summary

The effect of heterogeneity of residual variance on the efficiency of estimation of regression coefficients is examined in a practical way. Heterogeneity is shown to affect efficiency adversely to a degree much greater than has been commonly supposed, so much so, indeed, that raw data for regression analysis should always be systematically tested for residual variance homogeneity before processing. Suggestions are made for ensuring greater efficiency in this regard. Rates of one kind or another should always be used in regression in preference to absolute figures. Adjustment of data prior to regression may be fairly rough-and-ready since it is shown that a small degree of residual heterogeneity is not significantly inimical to efficiency.

In regression we are concerned with deriving relationships or laws from a matrix of data. In the simplest case of two variables linearly related we write

$$(1) \qquad \bar{y} = \alpha + \beta x,$$

conceived as a kind of ideal relationship between many pairs of associated variables, time series or other, (x_t, y_t) for $t = 1, 2, \ldots, T$. In regression relationship, as distinct from "associative" relationship [1], the x is regarded as measured without error, i.e. its values are those of the given series x_t, for calculation purposes, though the law (1) may be presumed to apply to any value of x. $\bar{y}$ does not, however, coincide with the observed y_t even when $x = x_t$. It differs from it by the error term u_t so that

$$(2) \qquad y_t = \bar{y}_t + u_t.$$

Workers in applied statistics almost invariably regard the error term u_t as an intrusive element, an unfortunate necessity. Knowing little or nothing about u_t we attribute to it such properties, usually stochastic, as will enable us to derive estimates of the coefficients α and β in (1) by the simplest methods.

The simplest assumption is that u_t is a random sample from $N(0, \sigma^2)$ where σ^2 is unknown but may be estimated from the data. Note that on an indefinitely large number of replications of the experiment of which our data represent merely one such, the error variance σ^2 is presumed the same for all x_t. Such an hypothesis may, in certain cases, be quite untenable, and demonstrably so. For instance, suppose that one is studying the relationship between average number of cattle and farm size. Suppose also that, as in the example studied later, average number of cattle on 8-acre farms is 4.5 and on 73-acre farms 27.4. It is quite ridiculous to suppose that the variance σ^2 on the smaller farms is the same as on the larger farms; it would be much more plausible to assume that variance increased regularly with size of farm, as will be found in the example to be the case. It would even be quite easy to derive the relationship. In general this relationship will not be known. It will presently be shown that estimates may be made of α and β in (1) by having initial regard to some relationship between variance and regressor,more accurate than by least square (LS) procedure applied to the raw data.

General ML case

The classicial procedure of solution, i.e. the estimation of α and β in the model,

$$(3) \qquad y_t = \alpha + \beta x_t + u_t,$$

where u_t is $N(0, \sigma^2)$ is by LS which is equivalent to maximum likelihood (ML) when u, as is postulated, is normal with σ^2 the same for each observation; an estimate s^2 of σ^2 also emerges from the procedure. It is natural to enquire if, by ML, a solution can be found by assuming that the tth observation error has the distribution $N(0, \sigma_t^2)$ where the σ_t^2 are also to be estimated from the exercise. The log frequency z is then

(4) $$z = -\frac{T}{2}\log 2\Pi - \Sigma\log\sigma_t - \tfrac{1}{2}\Sigma(y_t - \alpha - \beta x_t)^2/\sigma_t^2$$

By partial differentiation,

(5) $$\frac{\delta z}{\delta\sigma_t} = -\frac{1}{\sigma_t} + \frac{(y_t - \alpha - \beta x_t)^2}{\sigma_t^3} = 0,$$

giving

(6) $$\sigma_t^2 = (y_t - \alpha - \beta x_t)^2.$$

Also

$$\frac{\delta z}{\delta\alpha} = \Sigma(y_t - \alpha - \beta x_t)/\sigma_t^2$$

(7) $$= \Sigma(y_t - \alpha - \beta x_t)^{-1}\text{(from (6))} = 0;$$

Similarly,

(8) $$\frac{\delta z}{\delta\beta} = \Sigma x_t(y_t - \alpha - \beta x_t)^{-1} = 0.$$

Estimates of the coefficients α and β as a and b are, in theory, determinable from (7) and (8). Of course, there is no possibility of an exact solution, as in the classical case. They can be solved by 2-dimensional iteration. A first approximation would be found as a_o and b_o by classical least squares. The values of $\left(\frac{\delta z}{\delta\alpha}\right)_o$ and $\left(\frac{\delta z}{\delta\beta}\right)_o$ for these values are calculated. If these be small, as one would hope, values would be found for a grid of (a, b) with values near to these. The approximate answer would be found by inverse interpolation. An obvious difficulty will be the aberrations which will occur for some values of t for which $(y_t - a - bx_t)$ are very small, which may yield a dominatingly large value of its reciprocal. As will appear, this method is not suggested aspracticable.

A specific ML and LS comparison.

It may never be necessary to have recourse to the foregoing procedure since there may be other easier ways of obtaining nearly as accurate results. The guiding principle must always be to ensure that, in magnitude, the random variance is constant, or may be plausibly regarded as near constant, for all sets of observations. Suppose the model is

$$(9) \qquad y_t = \alpha + \beta x_t + \lambda_t u_t,$$

where u_t is $N(O, \sigma^2)$, as before, and the λ_t known,or estimable. The method proposed is to obtain ML estimates of α and β by converting (9) into a bivariate regression problem by dividing across by the known λ_t and solving in the usual way by LS. The procedure will now be applied by using specific values for λ_t. The object of the exercise will be to determine if concern about what we shall term residual heterovariance is of any practical importance and, if it is, to suggest ways of dealing with it. To study this question realistically the most useful model which the writer has succeeded in evolving is the following :-

$$(10) \qquad y_t = \alpha + \beta x_t + x_t^{\lambda} u_t, \quad t = 1, 2, \ldots, T,$$

where x_t and y_t, the observations, are assumed positive. λ is a know constant $\geqslant 0$ and u_t is $N(0,\sigma^2)$. It will be noted that model (10) postulates regular increase (in greater or lesser degree according to the value of λ) of residual variance with the magnitude of x_t. Dividing (10) across by x_t^{λ},

$$(11) \qquad z_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_t,$$

where $z_t = y_t x_t^{-\lambda}$, $x_{1t} = x_t^{-\lambda}$, $x_{2t} = x_t^{1-\lambda}$, $\beta_1 = \alpha$, $\beta_2 = \beta$. The ML = LS solution of (11) in matrix form is the familiar

$$(12) \qquad \begin{cases} b = (zX')(XX')^{-1} = \beta + (uX')(XX')^{-1}, \\ \text{Var-Covar}(b) = \sigma^2 (XX')^{-1}, \end{cases}$$

where, in this case,

$$(13) \qquad XX' = \begin{vmatrix} S(-2\lambda) & S(1-2\lambda) \\ S(1-2\lambda) & S(2-2\lambda) \end{vmatrix},$$

the sums S being given as

$$(14) \qquad S(\kappa) = \sum_1^T x_t^{\kappa}.$$

Here we shall be interested only in b_2, the estimate of $\beta_2 = \beta$. Its variance from the second of (12) is

$$(15) \qquad \text{Var}(b_2) = \sigma^2 S(-2\lambda)/[S(-2\lambda)S(2-2\lambda) - S^2(1-2\lambda)].$$

For straight LS treatment (10) is taken in the form

$$(16) \qquad y_t = \alpha + \beta x_t + v_t,$$

where $v_t = x_t^{\lambda}u_t$, whence the estimate b of β and its variance are

$$(17) \qquad \begin{aligned} b &= \Sigma\, y_t(x_t - \bar{x})/\Sigma(x_t - \bar{x})^2 \\ &= \beta + \Sigma\, v_t(x_t - \bar{x})/\Sigma(x_t - \bar{x})^2 \end{aligned}$$

$$\mathrm{Var}(b) = \sigma^2\Sigma\, x_t^{2\lambda}(x_t^2 - 2x_t\bar{x} + \bar{x}^2)/[\Sigma(x_t - \bar{x})^2]^2$$

$$(18) \qquad = \sigma^2[S(2\lambda + 2) - 2\bar{x}\; S(2\lambda + 1) + \bar{x}^2 S(2\lambda)]/[\Sigma(x_t - \bar{x})^2]^2.$$

If the x_t are all ordinary magnitudes, (18) shows that var(b) is $O(T^{-1})$; hence, though the solution (17) is not ML, the estimate b is consistent. Since the estimate b_2 is ML we know that, for $\lambda > 0$, var(b_2), given by (15) < var(b), given by (18). The answer to the question posed above involves comparison of the respective values. Unfortunately it is not possible to make any general assessment, i.e. for any positive vector x, closer than the two variance formulae. Instead we shall have to construct a numerical example, deemed illustrative of various realistic circumstances.

Scale. Both ML and LS estimates of β , represented respectively by b_2 and b, become b_2/p and b/p respectively when all the x_t are multiplied by a constant p, as will appear from the first of (12) and (17). Similarly, the variance formulae (12) (second) and (18) show that such transformation yields the same multiplier, namely $p^{2\lambda-2}$ for both var(b_2) and var(b). Here, at least, are points of similarity.

Example

The vector x is the set of natural numbers (1, 2, ... T), illustrative of the ~~general~~ case of equally spaced regressors. It is easy to show, from (12) and (18) that both estimates are consistent for $\lambda < 1.5$. However, for $\lambda = 1.5$, b_2 is still consistent (though variance is only $O(\log^{-1}T)$ but b_2 is not. So large a value as 1.5 for λ may be regarded as unrealistic; we shall concern ourselves merely with the range $1 \geq \lambda \geq 0$. Of course, for $\lambda = 0$ the estimates are identical. Taking $\sigma^2 = 1$ for the variance and using certain values of $T \leq 50$, the results are shown in the accompanying table.

Read vertically the table shows, for each value of T, that the efficiency of the LS estimate b dwindles sharply with increased residual heterogeneity. When λ is small, on the other hand, i.e. when the tendency for residual variance to increase with x is slight, the loss of efficiency by use of LS on the raw data is inconsiderable: we infer that for practical purposes it will suffice if the elimination of hetero-variance, discussed in the next section, is not complete, i.e. that the rough-and-ready methods proposed may be used with effect for elimination.

Read horizontally there is for each λ-value a tendency for relative efficiency to decline, though the effect gets less marked as T increases.

For the purpose of the paper it will not be necessary to consider a wide range of regressor types x

Comparison of Values of Residual Variance and Relative Asymptotic Efficiency for Maximum Likelihood (b_2) and Least Squares (b) Estimates of the Regression Coefficient when x is $\{1, 2, \ldots, T\}$.

λ	Variance and efficiency ratio	T				
		10	20	30	40	50
	Var(b_2)	.01212	$.0^{2}1504$	$.0^{3}4449$	$.0^{3}1876$	$.0^{4}9604$
0	Var(b)	"	"	"	"	"
	Ratio	1	1	1	1	1
	Var(b_2)	.01583	$.0^{2}2190$	$.0^{3}6936$	$.0^{3}3076$	$.0^{3}1638$
0.1	Var(b)	.01611	$.0^{2}2245$	$.0^{3}7143$	$.0^{3}3177$	$.0^{3}1695$
	Ratio	.9826	.9755	.9710	.9682	.9664
	Var(b_2)	.02072	$.0^{2}3186$	$.0^{2}1078$	$.0^{3}5017$	$.0^{3}2777$
0.2	Var(b)	.02207	$.0^{2}3498$	$.0^{2}1202$	$.0^{3}5650$	$.0^{3}3146$
	Ratio	.9388	.9108	.8968	.8880	.8827
	Var(b_2)	.04794	.01012	$.0^{2}4171$	$.0^{2}2242$	$.0^{2}1390$
0.5	Var(b)	.06667	.01579	$.0^{2}6897$	$.0^{2}3846$	$.0^{2}2449$
	Ratio	.7191	.6409	.6047	.5829	.5676
	Var(b_2)	.22399	.08410	.04975	.03484	.02664
1.0	Var(b)	.54424	.25549	.16681	.12381	.09843
	Ratio	.4116	.3292	.2982	.2814	.2706

or different modes of residual heterovariance. The table shows plainly that when heterovariance takes the form of regular increase with the regressor, inefficiency of estimate must be suspected, tested for and, if necessary, eliminated. The regressor vector chosen does not, however, cover the general case of equi-spaced values of positive x_t. One may infer, in fact, that the range ratio between 1 and the various values of T(i.e. T:1 = T) adversely affects efficiency. To show what happens when the range ratio is small, given T, we contrast the efficiency for T = 20, λ = 0.5 for (1) $x = \{1, 2, \ldots, 20\}$ given in the table for (2) $x = \{31, 32, \ldots 50\}$, with range ratios 20 and 50 : 31 = 1.6 respectively :-

	$Var(b_2)$	Var(b)	Efficiency ratio
(1)	.01012	.01579	.6409
(2)	.05991	.06090	.9837

These figures convey a more than broad hint that heterovariance does not adversely affect LS estimates based on raw data when the range ratio is low, though, of course, the degree of residual heterogeneity is also reduced in (2) by the model adopted (10).

In a paper [2] published some years ago, the writer analysed agricultural data on farms of different sizes in an Irish county. It was, of course, found that all classes of statistics (livestock, crops and output) increased per farm very regularly with farm size; of greater interest was the precession on farms of different size, of the various statistics studied per 100 acres and per person engaged. The coefficient of variation was also analyzed, showing, it is true, in all cases a tendency to decline with increasing farm size. The standard deviation for farm units, however, increased regularly with farm size: for instance, for total cattle it was 4.7 for farms of 6 - 10 acres and 14.24 for farms 71 - 75 acres. It is calculated that in (10) $\lambda = 0.53$ approximately for cattle on farm units. If the "realization" were one farm in each of 14 size classes, i.e. 14 farms in all, regression of number of cattle on farm size using all the raw data would have an efficiency of only about 73% as compared with the accuracy of estimation obtainable if the data about the precession of variance were obtainable and taken into account.* How this can be done on the usual single realization will be indicated in the next section.

Conclusion

This paper owes its inception to the writer's wish to clear up (for his own information and with no thought of publication) a small point of theory, exemplified in the simplest possible case he could devise. Of course, in very general conditions, the ML solution is asymptotically the most efficient, and is also, in practice, usually the most efficient for samples of ordinary size. Before the present results became available, like most other statisticians, he did not attach much practical importance to the hypotheses for the residual error, normality with population mean zero and homovariance, the latter meaning that for every regressor the population error variance is the same. (We are not concerned here with residual

*Found from (15) and (18) with x (14 elements, equally spaced, in acreage) = $\{8, 13, 18, \ldots, 73\}$. With $\lambda = 0.53$, $\sigma^2 = 1$, variances were var(b) = .0090295, var(b_2) = .0065774.

autoregression). The presumption was that these hypotheses, necessary for the LS determination of the coefficients, did not matter much, that a marked degree of residual heterovariance might not effect the efficiency of coefficient variance very markedly. The foregoing results show this anticipation was not correct. When the residual variance increases more or less regularly with the magnitude of the regressor LS estimation applied to the raw data in the usual way may yield estimates which may be only 50 per cent as efficient as if the data were homovariant. Such a statement means broadly that we would be in the situation of rejecting 50 per cent of our data, usually hard to come by, in making our estimates. The writer freely admits that residual variance may vary in other ways, which may not be so inimical to estimation efficiency. Nonetheless, in his opinion the case he has considered, namely regular variation, is that most likely to be encountered in practice, that it should always be suspected when the range of variation in the regressor is wide and that steps should be taken to counteract it. In this paper the problem of simple regression (i.e. one regressor) only is dealt with. Obviously, the findings would apply also to multivariate regression.

As a matter of routine before embarking on a regression we should take steps to ensure that the hypothesis of approximate residual homovariance is plausible. Assuming the original data positive it should be used without adjustment only when the range of variation (say ratio value of largest to smallest) is small. If at all practicable, rates of one kind or another should be used in preference to raw data. The common course of introducing a scale variable (e.g. using absolute number of births and absolute size of population, instead of the single variate birthrate) is not to be recommended, though we may be bemused by the inevitably large value of R^2 with such treatment. Our concern should be, as in elementary statistics, to compare like with like, even standardized rates to crude rates. Apart from efficiency of estimation, this is merely commonsense; in simple regression we examine, on a cause-effect basis two measured phenomena; we should try to equalize (or elminate) other possibly relevant sources of variability from the comparison. It usually happens in statistical practice that the sensible course is also efficient. In the agricultural example cited earlier it would obviously have been more appropriate and informative to regress the statistic cows per 100 acres rather than cows per farm on farm size; everyone would expect there to be more cows on farms of 100 acres than on farms of 10 acres; as we now know the per 100 acre approach will also yield much more accurate estimates of the parameters involved.

We now consider treatment of the raw data to eliminate residual heterovariance. If the range of values remains wide and if the number of sets of observations is reasonably large, these might be deivided into, say, five or six groups according to the magnitude of the regressor (or to the magnitude of the principal component of the regressors in the multivariate case: this treatment would be designed to take account of non-linearity of regressors.

The estimated residual variance s_i^2 (i = 1, 2, ... 5 or 6) would be calculated for each group. If this seems to vary regularly with $\bar{x}_i$ (the group mean of the regressor values) strike a rough regression of log s_i on log $\bar{x}_i$; so the positive value of λ would be determined and the model in simple regression would be $y_t = \alpha + \beta_\lambda x_t + x_i^\lambda u_t$. Divide through by x_i^λ and apply bivariate LS to estimate α and β. Even if the procedure be very rough indeed it would appear that the resulting estimates will be much more accurate than if LS were applied to a model $y_t = \alpha + \beta x_t + v_t$.

The paper raises a problem of exegesis. It would appear, at first sight, than an exposé of simple regression cannot remain matter for elementary statistical manuals, as it has been heretofore, since the student must, at a very early stage become aware of the statistical notions of consistency and efficiency usually regarded as "advanced". This need not be the case if the commonsensible approach of a previous paragraph, with its emphasis on rates, be adopted. There does not, however, seem to be any way of avoiding treatment of bivariate regression by least squares at an earlier stage than is at present usual.

REFERENCES

[1] Geary, R. C., Some Remarks about Relations between Stochastic Variables: A Discussion Document. Review of The International Statistical Institute, Vol 31:2, 1963.

[2] Geary, R.C., Variability in Agricultural Statistics on Small and Medium-Sized Farms in An Irish County. Journal of the Statistical and Social Inquiry Society of Ireland, Vol XIX, 1956-57