The Efficiency of Taking First Differences in Regression Analysis: A Note

J. A. TILLMAN

In a recent article, Geary [1972] discussed the merit of taking first differences to deal with the problems that trends in data present in regression analysis. Geary gave examples of situations where this procedure leads to highly inefficient estimates of the regression coefficients. The first difference transformation has also been suggested as an appropriate way of dealing with multicollinearity among the independent variables, for example Kane [1968, p. 280]. This note generalises Geary's results and shows that transforming the data by taking first differences for such a purpose cannot improve the efficiency of the regression estimates but will in general cause a reduction in efficiency. Although the usual least squares formula is not appropriate for calculating the variance of regression estimates obtained from the transformed data, they are often used for this purpose. Their relation to the variance of the regression estimates obtained from the untransformed data is examined.

For completeness this note concludes with a brief discussion of first differencing to deal with serially correlated disturbances in the regression model.

Trend and Multicollinearity

Let the regression model be

$$y = X\beta + u \tag{1}$$

where y is a $n \times 1$ vector, X a $n \times k$ matrix of rank k, and β a vector of k parameters to be estimated. u is an unobserved $n \times 1$ vector of disturbances such that $E(u) = 0$, $E(u\,u') = \sigma^2 I$. For such a regression model the best linear unbiased estimator is the familiar least squares estimator

$$b = (X'X)^{-1}X'y$$

with covariance $\text{Cov}(b) = \sigma^2(X'X)^{-1}$

$$= \sigma^2 \Sigma_b$$

However, trends in the data may give rise to implausibly large r values. Another problem frequently met with is that of a high degree of multicollinearity between the independent variables. This may result in the estimated coefficients having insignificant t ratios. Taking first differences is sometimes employed as a way out of these difficulties. We will show that this cannot lead to more efficient estimates if the underlying regression is given by (1).

Premultiplying (1) by the $n-1 \times n$ matrix T transforms the regression to first differences, thus

$$Ty = TX\beta + Tu \tag{2}$$

$$\text{where} \quad T = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 & 0 \\ & \vdots & & & & \vdots & \\ 0 & 0 & 0 & & -1 & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{bmatrix}$$

If an intercept term is included in the regression, first differencing reduces this to zero. To avoid this, both the dependent and independent variables are expressed as deviations about their means. This eliminates the need for an explicit intercept term in (1).

Let $y^\star = Ty$, $X^\star = TX$ and $u^\star = Tu$. Hence $y^\star$, $X^\star$ and $u^\star$ have $n-1$ rows and $E(u^\star u^{\star\prime}) = \sigma^2 TT'$. Therefore (2) may be re-written as

$$y^\star = X^\star\beta + u^\star \tag{3}$$

where β is now to be estimated from (3). But, to quote Geary, "Invariably, however, when the regression problem is deltaised the assumption is made that the error term '$u^\star$' is regular, which assumption amounts to a wrong specification if the basic model is (1)." The least squares estimator using the deltaised data $y^\star$, $X^\star$ is

$$b^\star = (X^{\star\prime}X^\star)^{-1}X^{\star\prime}y^\star$$

with covariance

$$\text{Cov}(b^\star) = \sigma^2(X^{\star\prime}X^\star)^{-1}X^{\star\prime}TT'X^\star(X^{\star\prime}X^\star)^{-1}$$

$$= \sigma^2\Sigma_{b_\star}$$

whereas the best linear unbiased estimator is now

$$\hat{\beta}^\star = (X^{\star\prime}(TT')^{-1}X^\star)^{-1}X^{\star\prime}(TT')^{-1}y^\star$$

with covariance

$$\mathrm{Cov}(\hat{\beta}^\star) = \sigma^2(X^{\star\prime}(TT')^{-1}X^\star)^{-1}$$
$$= \sigma^2\Sigma_{\hat{\beta}^\star}$$

The question to be resolved is that of the relative efficiency among the three estimators b, $b^\star$ and $\hat{\beta}^\star$. Successively applying the Generalised Gauss Markov theorem yields the inequality[1]

$$\mathrm{Cov}(b) \leqslant \mathrm{Cov}(\beta^\star) \leqslant \mathrm{Cov}(b^\star)$$

Thus $b^\star$ is the least efficient of the three estimates. This loss of efficiency is due both to incorrectly assuming that $E(u^\star u^{\star\prime}) = \sigma^2 I$ and to "losing" one observation in taking first differences. If T were a $n \times n$ matrix of full rank $\hat{\beta}^\star$ would be identical to b, but $b^\star$ would still, in general differ from b.

If, after taking first differences to remove trend or multicollinearity, the $u^\star$ are regarded as being regular then not only will $b^\star$ be used instead of $\hat{\beta}^\star$ but the covariance of $b^\star$ is estimated (incorrectly) by

$$\hat{\mathrm{C}}\mathrm{ov}(b^\star) = \sigma^2(X^{\star\prime}X^\star)^{-1}$$

First differencing would be judged as having successfully dealt with multicollinearity if $\mathrm{Cov}(b^\star)$ is smaller than $\mathrm{Cov}(b)$. It is therefore of some interest to investigate the extent to which such a reduction is possible. We shall obtain bounds for the ratio of the generalised covariances, $|\hat{\mathrm{C}}\mathrm{ov}(b^\star)| / |\mathrm{Cov}(b)|$.

The matrix of regression vectors X may be written

$$\underset{nk}{X} = \underset{nk}{P}\,\underset{kk}{K}$$

where $P'P = I$ and K is non-singular.
Hence

$$|X'X| = |K'K|$$
$$= |K|^2$$

and

$$|X^{\star\prime}X^\star| = |X'T'TX|$$
$$= |K'P'T'T'TPK|$$
$$= |P'T'TP|\;|K|^2$$
$$= |X'X| \prod_{i=1}^{k} Ch_i(P'T'TP)$$

1. This is discussed further in Appendix 1.

where we have made use of the fact that the determinant of a symmetric matrix is equal to the product of its characteristic roots. The ith largest characteristic root of a symmetric matrix A may be denoted by $Ch_i(A)$.

Although the characteristic roots of $P'T'TP$ depend on X, Cauchy's inequality enables them to be bounded by the characteristic roots of $T'T$.

Cauchy's Inequality

If A is an $n \times n$ symmetric matrix and R an $n \times k$ matrix such that $R'R = I$,

$$Ch_{i+n-k}(A) \leqslant Ch_i(R'AR) \leqslant Ch_i(A) \qquad i = 1 \ldots k$$

The next step is to evaluate the characteristic roots of $T'T$.

But

$$T'T = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & & 0 & 0 & 0 \\ & \vdots & & & & \vdots & \\ 0 & 0 & 0 & & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 1 \end{bmatrix}$$

It is known, for example see Anderson [1971], that

$$Ch_i(T'T) = 2\left(1 - \text{Cos}\frac{(n-i)\pi}{n}\right)$$

Therefore

$$4 > Ch_1(T'T) > \ldots > Ch_n(T'T) = 0$$

The characteristic vector of T' corresponding to $Ch_n(T'T)$ is a vector of identical elements orthogonal to the columns of X. Therefore Cauchy's inequality contracts to give

$$Ch_{i+n-k-1}(T'T) \leqslant Ch_i(P'T'TP) \leqslant Ch_i(T'T) \qquad i = 1 \ldots k \quad (4)$$

The upper bound is attained if the columns of P correspond to the characteristic vectors of $T'T$ giving the k largest roots. These vectors may be written as P_2. The lower bound is attained if the columns of P correspond to vectors that yield the k smallest roots excluding $Ch_n(T'T)$. These may be written as P_1.

After some simplification we obtain bounds for the ratio of the generalised covariances,

$$\frac{1}{\prod_{i=1}^{k} Ch_i(T'T)} \leqslant \frac{|\,\hat{\text{C}}\text{ov}(b^*)\,|}{|\,\text{Cov}(b)\,|} \leqslant \frac{1}{\prod_{i=1}^{k} Ch_{i+n-k-1}(T'T)} \qquad (5)$$

The upper bound is attained if $X=P_1K$, the lower if $X=P_2K$. The bounds are tabulated for various values of k and n in Table 1. The lower bound approaches $\frac{1}{4k}$ as n increases but the upper bound has no limit as n increases.

The conditions for the ratio of the generalised covariances to attain its maximum value has an interesting economic interpretation. In many economic applications, particularly those using time series data the regression vectors are slowly changing, that is the change from one period to the next is small in relation to the total change over n periods. In such a case the matrix of regression vectors X is approximately equal to P_1K. It might be conjectured that for small divergences from $X=P_1K$ the relation of $|\hat{C}ov(b)^*|$ to $|Cov(b)|$ is given approximately by the upper bound in Table 1. Therefore in many economic applications the incorrect estimate $\hat{C}ov(b^*)$ will be larger than the covariance of the original b. Thus even if the original estimator does not suffer from multicollinearity the incorrect estimate b^* will appear to do so.

TABLE 1

$$\frac{|\hat{C}ov(b^*)|}{|Cov(b^*)|}$$

			k			
			1	2	3	4
n	10	Upper bound	10·2	26·7	32·4	23·5
		Lower bound	0·250	0·071	0·022	0·009
	30	Upper bound	$1·6\times 10^2$	$6·5\times 10^3$	$1·2\times 10^5$	$1·2\times 10^6$
		Lower bound	0·250	0·063	0·016	0·004

k = number of independent variables, k does not include the constant term.

For n = 30, the figures for the Upper Bound have been rounded.

First Order Serial Correlation

For completeness we conclude by briefly describing another use for which the first difference transformation has been proposed. In many applications of the regression model, particularly those using time series data it is suspected that the disturbance term u follows a first order autoregressive process. That is

$$u_t = \rho u_{t-1} + \varepsilon_t$$

where the ε_t are independent identically distributed error terms.

Premultiplying equation (1) by the $n-1 \times n$ matrix Q corresponds to the generalised first difference transformation and regularises the error term. Thus

$$Qy = QX\beta + Qu \tag{6}$$

where

$$Q = \begin{bmatrix} -\rho & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -\rho & 1 & & 0 & 0 & 0 \\ & \vdots & & & & \vdots & \\ 0 & 0 & 0 & & -\rho & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & -\rho & 1 \end{bmatrix}$$

It may easily be shown that $E(Q\,u\,u'Q')=\sigma^2 I_{n-1}$. Unlike the problem of trend or multicollinearity, the purpose of the transformation is now to deal with the irregular error term.

Kadiyala [1968] discussed the efficiency of the least squares estimator obtained using the transformed data Qy and QX relative to the ordinary least squares estimator. For the special case of X being a column of ones the estimator obtained from (6) was always less efficient than the least squares estimator and as ρ approached one the efficiency dropped to zero. It is for the values of ρ close to one that the first difference transformation has been recommended for dealing with first order serial correlation. A better procedure is to use a transformation $Q^\star$, where $Q^\star = \begin{pmatrix} q \\ Q \end{pmatrix}$ and $q=(\sqrt{1-\rho^2}, 0, \dots, 0)$. If ρ is unknown it can be estimated from the ordinary least squares residuals, for example Johnston [1972].

Conclusion

The first difference transformation is not an appropriate way of dealing with trend on multicollinearity in regression analysis. Transforming the data in this way cannot increase the efficiency of the regression estimates and will in general reduce the efficiency.

For first order serial correlation the transformation $Q^\star$ is superior to the generalised first difference transformation Q.

I would like to thank T. Muench for his helpful comments on an earlier draft of this paper.

University of Massachusetts, Boston.

REFERENCES

[1] Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, J. Wiley, New York.
[2] Geary, R. C. (1972), "Two Exercises in Simple Regression", *Economic and Social Review*, Vol. 3, No. 4, July, 1972.
[3] Johnston, J. (1972), *Econometric Methods*, 2nd edition, McGraw Hill, New York.
[4] Kadiyala, K. R. (1968), "A Transformation Used to Circumvent the Problem of Autocorrelation", *Econometrica*, 36, 93–96.
[5] Kane, E. J. (1968), *Economic Statistics and Econometrics*, Harper and Row.

APPENDIX 1

The $k \times k$ covariance matrics of b, $\hat{\beta}^\star$ and $b^\star$ are given by $\sigma^2\sum_b$, $\sigma^2\sum_{\hat{\beta}^\star}$ and $\sigma^2\sum_{b^\star}$ respectively. The inequality between the covariances is to be understood in the following way. If A and B are symmetric positive definite matrices of the same order then $A \leqslant B$, if and only if $A-B$ is positive semi-definite.

The Generalised Gauss Markov theorem states that for the regression model (1), b has minimum variance among all unbiased linear estimators of β. Since $\hat{\beta}^\star$ is another unbiased linear estimator of β it follows that $\mathrm{Cov}(b) \leqslant \mathrm{Cov}(\hat{\beta}^\star)$.

Similarly for model (3), $\hat{\beta}^\star$ has minimum variance among all unbiased linear estimator of β where the data is of the form $y^\star$, $X^\star$. But $b^\star$ is also an unbiased linear estimator of β hence

$$\mathrm{Cov}(\hat{\beta}^\star) \leqslant \mathrm{Cov}(b^\star)$$