The Pakistan Development Review
Vol. XXVII, No. 4 Part II (Winter 1988)

Combining Yearly and Quarterly Data in Regression Analysis

EATZAZ AHMAD*

I. INTRODUCTION

Data deficiency is often a problem in regression analysis. The problem, for example, may be due to non-availability of data on some variable, missing observations, lack of information due to multicollinearity and measurement errors, etc. Various approaches have been suggested to deal with the problem depending on its precise nature. One such problem we want to focus our attention on is the lack of time disaggregated data in time-series regression analysis. In particular, observations on some variables over a shorter time interval like a quarter may be limited in number while the corresponding observations over a longer time interval like a year are available for a long period of time. The number of quarterly observations may not be sufficient to estimate the desired relationship with acceptable degrees of freedom. On the other hand, estimation with yearly data may require the use of a long time series going way back into the past. The estimates thus obtained may not capture the relationship prevailing at present or in the recent past and, therefore, mislead the researcher. In addition, the use of yearly data may also result in lack of degrees of freedom.

One possible approach to deal with the problem is to convert the yearly data into quarterly data by using some distribution scheme and combine with these data the other quarterly observations available. Friedman (1962) suggests various non-correlation methods of distributing a time aggregated series into a time disaggregated series. Chow and Lin (1971), Friedman (1962), Hsiao (1979) and Palm and Nijman (1982) also suggest various methods of using a time disaggregated related series to distribute a time aggregated series over shorter time intervals.

The problem with any method of distribution is that it introduces measurement error and related problems in regression analysis. Due to this reason we suggest another approach which is quite simple and does not involve distribution of yearly data over quarterly intervals. One can simply pool quarterly and yearly data adjusting for heteroscedasticity introduced by the pooling. It is shown that the

*The author is Assistant Professor in the Department of Economics, Quaid-i-Azam University, Islamabad. He is grateful to Professor F. T. Denton of McMaster University for his comments on an earlier draft of this paper.

estimators of regression coefficients with pooled data have smaller variances as compared to the estimators with yearly or quarterly data.

The paper is organized as follows. In Section 2 we present the model and derive an Ordinary Least Squares estimator and the corresponding variance-covariance matrix of the vector of regression coefficients with alternative data sets. Relative efficiency of the alternative estimators is compared in Section 3. Section 4 presents the conclusions of the paper with some thoughts on future research.

2. THE MODEL AND ITS ESTIMATION WITH ALTERNATIVE DATA SETS

Let the relationship to be estimated be described by the following linear regression equation:

$$y_{ti} = b_1 + b_2 x_{2ti} + \ldots + b_k x_{kti} + u_{ti} \quad \ldots \quad \ldots \quad \ldots \quad (1)$$

where, t and i refer to year and quarter respectively. We assume that:

(1) All the variables are flow variables;
(2) There is no lagged variable in the equation;
(3) All the x variables are non-stochastic;
(4) u_{ti} is randomly distributed with $E(u_{ti}) = 0$ for all t and i; and
(5) $E(u_{ti}\ u_{sj}) = \sigma^2$ for all $t = s$ and $i = j$
$= 0$ for all $t \neq s$ or $i \neq j$.

Owing to assumptions (1) and (2) we can conveniently define yearly observations as:

$$Y_t = \sum_{i=1}^{4} y_{ti} \text{ for all } t$$

$$X_{jt} = \sum_{i=1}^{4} x_{jti} \text{ for all } t \text{ and } j$$

$$U_t = \sum_{i=1}^{4} u_{ti} \text{ for all } t.$$

Thus we can write Equation (1) for yearly observations as:

$$Y_t = 4b_1 + b_2 X_{2t} + \ldots + b_k X_{kt} + U_t \quad \ldots \quad \ldots \quad \ldots \quad (2)$$

where U_t satisfies the following obvious properties.

(4′) U_t is randomly distributed with $E(U_t) = 0$ for all t and

(5′) $E(U_t\ U_s) = 4\sigma^2$ for all $t = s$

$= 0$ for all $t \neq s$.

With no loss of generality, we assume that the quarterly observations are available over a whole number of years, say m. In addition, n yearly observations are also assumed to be available. As is more likely, it is assumed that the yearly observations appear before the quarterly ones. We consider the following three options available to estimate the relationship.

Option I

Use yearly data over $n + m$ years. With this option the regression model can be written as:

$$\begin{bmatrix} Y \\ Y^* \end{bmatrix} = \begin{bmatrix} X \\ X^* \end{bmatrix} B + \begin{bmatrix} U \\ U^* \end{bmatrix}$$

where,

$$Y = \begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{bmatrix}, \quad Y^* = \begin{bmatrix} Y_{n+1} \\ . \\ . \\ . \\ Y_{n+m} \end{bmatrix}, \quad U = \begin{bmatrix} U_1 \\ . \\ . \\ . \\ U_n \end{bmatrix}, \quad U^* = \begin{bmatrix} U_{n+1} \\ . \\ . \\ . \\ U_{n+m} \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ . \\ . \\ . \\ b_k \end{bmatrix}$$

$$X = \begin{bmatrix} 4 & X_{2,1} & \ldots & X_{k,1} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 4 & X_{2,n} & \ldots & X_{k,n} \end{bmatrix} \text{ and } X^* = \begin{bmatrix} 4 & X_{2,n+1} & \ldots & X_{k,n+1} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 4 & X_{2,n+m} & \ldots & X_{k,n+m} \end{bmatrix}$$

The best linear estimator of B, its mean vector and variance-covariance matrix are as follows.[1]

$$B_I = (X'X + X'_*X_*)^{-1}\ (X'Y + X'_*Y_*)$$

$$E(B_I) = B$$

$$V(B_I) = 4\sigma^2\ \ (X'X + X'_*X_*)^{-1} \quad \ldots \quad \ldots \quad (3)$$

[1]It is assumed that the matrix $[X'\ X'_*]$ has full column rank k which is less than $n + m$.

Option II

Use only quarterly data over $4m$ quarterly observations. In this case the regression model can be written as:

$$y_* = x_* B + u_*$$

where,

$$y_* = \begin{bmatrix} y_{n+1,1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n+m,4} \end{bmatrix}, \quad u_* = \begin{bmatrix} u_{n+1,1} \\ \cdot \\ \cdot \\ \cdot \\ u_{n+m,4} \end{bmatrix}, \quad x_* = \begin{bmatrix} 1 & x_{2,n+1,1} & \cdots & x_{k,n+1,1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & x_{2,n+m,4} & \cdots & x_{k,n+m,4} \end{bmatrix}$$

and B is the same as defined before.

The best linear estimator of B, its mean vector and variance-covariance matrix under this option are:[2]

$$B_{II} = (x'_* x_*)^{-1} (x'_* y_*)$$

$$E(B_{II}) = B$$

$$V(B_{II}) = \sigma^{2} (x'_* x_*)^{-1} \quad \ldots \quad \ldots \quad \ldots \quad (4)$$

Option III

Combine $4m$ quarterly and n yearly observations. Since variance of U_t is four times as high as the variance of u_{ti}, we have the problem of heteroscedasticity. With pre-adjustment for heteroscedasticity, the model becomes:

$$\begin{bmatrix} ½ & Y \\ & y_* \end{bmatrix} = \begin{bmatrix} ½ & X \\ & x_* \end{bmatrix} B + \begin{bmatrix} ½ & U \\ & u_* \end{bmatrix}$$

where, Y, y_*, X, x_*, U, u_* and B are the same as defined earlier. With this option the best linear estimator of B, its mean vector and variance covariance matrix are:

$$B_{III} = (¼ X' X + x'_* x_*)^{-1} (¼ X' Y + x'_* y_*)$$

[2] The matrix x_* is assumed to have full column rank $k < 4m$.

$$E(B_{III}) = B$$

$$V(B_{III}) = \sigma^{2} (¼ X' X + x'_* x_*)^{-1} \quad \ldots \quad \ldots \quad \ldots \quad (5)$$

3. RELATIVE EFFICIENCY OF THE ALTERNATIVE ESTIMATORS OF B

Since all the three estimators of B outlined above are unbiased, their relative efficiency depends only on relative variances. For the comparison of variances we will use the following theorem taken from Maddala (1977).

Theorem

If B is a positive definite matrix and $A - B$ is positive semidefinite then $B^{-1} - A^{-1}$ is positive semidefinite.

Proof

Since B and hence B^{-1} are positive definite and $A - B$ is positive semidefinite, the matrix $B^{-1}(A - B)$ is positive semidefinite. Therefore the Equation $|B^{-1}(A - B) - \nu I|$ has all roots $\nu \geqslant 0$. But the roots of this equation are the same as the roots of $|(A - B) - \nu B|$ or $|A - (1 + \nu)B|$ or $|B^{-1} - (1 + \nu)A^{-1}|$ or $|A(B^{-1} - A^{-1}) - \nu I|$. Thus all $\nu \geqslant 0$ implies that $A(B^{-1} - A^{-1})$ is positive semidefinite. Since B is positive definite and $A - B$ is positive semidefinite, sum of the two, that is A is positive definite and so is A^{-1}. Multiplying this positive definite matrix (A^{-1}) by the positive semidefinite matrix $A(B^{-1} - A^{-1})$ gives a positive semidefinite matrix $B^{-1} - A^{-1}$. This completes the proof.

Let us now apply this theorem to compare variances of various estimators of B. We will consider the variance of a linear combination of the elements of B_r $(r = I, II, III)$, namely $c' B_r$ where, c is a $k \times 1$ column vector of known constants.

$Var(c'B_{III})$ Versus $Var(c'B_I)$

Consider the following matrix.

$$(¼ X'X + x'_* x_*) - (¼ X'X + ¼ X'_* X_*) = x'_* x_* - ¼ X'_* X_*$$

The element in jth row and hth column of this matrix is:

$$\sum_{t=n+1}^{n+m} \sum_{i=1}^{4} x_{jti} x_{hti} - ¼ \sum_{t=n+1}^{n+m} \left(\sum_{i=1}^{4} x_{jti} \right) \left(\sum_{i=1}^{4} x_{hti} \right) =$$

$$\sum_{t=n+1}^{n+m} \sum_{i=1}^{4} (x_{jti} - \bar{x}_{jt})(x_{hti} - \bar{x}_{ht}) = \sum_{t=n+1}^{n+m} \sum_{i=1}^{4} z_{jti} z_{hti}$$

where, $z_{jti} = x_{jti} - \bar{x}_{jt}$, $z_{hti} = x_{hti} - \bar{x}_{ht}$, $\bar{x}_{jt} = \sum_{i=1}^{4} x_{jti}/4$

and $\bar{x}_{ht} = \sum_{i=1}^{4} x_{hti}/4$

The above discussion implies that the whole matrix $(¼\, X'X + x'_* x_*) - (¼\, X'X + ¼\, X'_* X_*) = x'_* x_* - ¼ X'_* X_*$ can be written as $Z'_* Z_*$ where,

$$Z_* = \begin{bmatrix} 0 & x_{2,n+1,1} - \bar{x}_{2,n+1} & \cdots & x_{k,n+1,1} - \bar{x}_{k,n+1} \\ 0 & x_{2,n+1,2} - \bar{x}_{2,n+1} & \cdots & x_{k,k+1,2} - \bar{x}_{k,n+1} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 0 & x_{2,n+m,4} - \bar{x}_{2,n+m} & \cdots & x_{k,n+m,4} - \bar{x}_{k,n+m} \end{bmatrix}$$

Clearly, $Z'_* Z_* = (¼\, X'\, X + x'_* x_*) - (¼\; X'\, X + ¼\, X'_* X_*)$ is positive semidefinite. In addition, $(¼\, X'\, X + ¼\, X'_* X_*)$ is positive definite. Therefore, according to the theorem, the matrix $4\,(X'\, X + X'_* X_*)^{-1} - (¼\, X'\, X + x'_* x_*)^{-1}$ is positive semidefinite.

$$4\,\sigma^2\; c'(X'\, X + X'_* X_*)^{-1} c \geqslant \sigma^2\; c'\,(¼\, X'\, X + x'_* x_*)^{-1} c.$$

Or, according to Equations (3) and (5)

$$var\,(c'\, B_I) \geqslant var\,(c'\, B_{III}).$$

Notice that, if $x_{jti} = \bar{x}_{jt}$, that is, if there is no variation across quarters within a year then $Z'_* Z_*$ is positive as well as negative semidefinite. In this case $var(c'B_I)$ is equal to $var\,(c'B_{III})$. This is precisely what one should expect. If within a year variation across quarters is zero then disaggregating yearly observations into quarterly observations does not provide any additional information.

Var $(c'\, B_{III})$ Versus *Var* $(c'\, B_I)$

Now consider the matrix $(¼\, X'\, X + x'_* x_*) - x'_* x_* = ¼\, X'\, X$ which is obviously positive semidefinite. Since, in addition, $x'_* x_*$ is positive definite, the above

theorem implies that the matrix $(x'_* x_*)^{-1} - (¼\, X'\, X + x'_* x_*)^{-1}$ is positive semidefinite.

Therefore we can write:

$$\sigma^2\; c'\,(x'_* x_*)^{-1} c \geqslant \sigma^2\; c'\,(¼\, X'\, X + x'_* x_*)^{-1} c$$

This implies, according to Equations (4) and (5), that

$$var\,(c'\, B_{II}) \geqslant var\,(c'\, B_{III}).$$

This result does not require any explanation. Addition of observations to a given data set improves efficiency of the estimators.

Illustration: One Explanatory Variable Case

Consider the case of one explanatory variable without intercept:

$$y_{ti} = b\, x_{ti} + u_{ti}$$

In this case the variance of b_r $(r = I, II, III)$ can be calculated as follows:

$$var\,(b_I) = \frac{\sigma^2}{¼ \sum_{t=1}^{n+m} \left(\sum_{i=1}^{4} x_{ti}\right)^2}$$

$$var\,(b_{II}) = \frac{\sigma^2}{\sum_{t=n+1}^{n+m} \sum_{i=1}^{4} (x_{ti})^2}$$

$$var\,(b_{III}) = \frac{\sigma^2}{¼ \sum_{t=1}^{n} \left(\sum_{i=1}^{4} x_{ti}\right)^2 + \sum_{t=n+1}^{n+m} \sum_{i=1}^{4} (x_{ti})^2}$$

Calling the denominator of b_t as D_r respectively for $r = I, II$ and III, we can show that:

$$D_{III} = D_I + \sum_{t=n+1}^{n+m} \sum_{i=1}^{4} (x_{ti} - \bar{x}_t)^2 \text{ and}$$

$$D_{III} = D_{II} + \sum_{t=1}^{n} \left(\sum_{i=1}^{4} x_{ti}\right)^2$$

This implies that:

$var\,(b_{III}) < var\,(b_I)$ unless $X_{ti} = \bar{x}_t$ for $t = n+1, \ldots, n+m$

$t = 1, \ldots, 4$

$var(b_{III}) < var(b_{II})$ unless $\sum_{i=1}^{4} x_{ti} = 0$ for $t = 1, \ldots, n$.

5. CONCLUDING REMARKS

Aggregation of quarterly data into yearly data for any part of the sample period results in loss of efficiency. Using quarterly data alone when additional yearly data are available also results in loss of efficiency. The best use of the limited data is to pool yearly observations with the quarterly observations with appropriate adjustment for heteroscedasticity. The result can be generalized to include seasonal effects in the regression equation. It can be shown that pooling yearly data to a given set of quarterly data also improves efficiency of seasonal effects although the yearly data alone are useless to estimate seasonal effects.[3]

The research can be extended to develop tests for autocorrelation of first or fourth order (these are the most likely orders of autocorrelation at quarterly level). Our preliminary research suggests that it is quite complicated to determine the order of autocorrelation with pooled data. Once the order of autocorrelation is determined, one can use various procedures to improve the asymptotic efficiency of the estimates.

REFERENCES

Chow, G. C., and A. I. Lin (1971). "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series". *Review of Economics and Statistics.* Vol. CIII, No. 4. pp. 372–375.

Friedman, M. (1962). "The Interpolation of Time Series by Related Series". *Journal of American Statistical Association.* Vol. 57, pp. 729–757.

Hsiao, C. (1979). "Linear Regression Using Both Temporally Aggregated and Temporally Disaggregated Data". *Journal of Econometrics.* Vol. 10, No. 2, pp. 243–252.

Maddala, G. S. (1979). *Econometrics.* Tokyo: McGraw-Hill, Inc.

Palm, F. C., and T. E. Nijman (1982). "Linear Regression Using Both Temporally Aggregated and Temporally Disaggregated Data", *Journal of Econometrics.* Vol. 19, pp. 333–343.

[3] This result has been proved in the paper originally presented in the general meeting. The paper had to be reduced due to the space constraint.

Comments on "Combining Yearly and Quarterly Data in Regression Analysis"

In this paper the author has proposed an alternative approach to deal with the issue of non-availability of comparable data in regression analysis. In particular, the paper deals with the problem where both quarterly as well as annual observations on some variables are available but the number of each type of observation i.e. quarterly limited as a result of which it is difficult to estimate the desired relationship, by either using quarterly or annual data with acceptable degrees of freedom. Various approaches have been suggested in the literature to deal with such problems. A common weakness of most of these approaches is that they introduce measurement errors. This paper suggests that instead of estimating the desired relationship by either using quarterly or annual data the researcher should simply pool the quarterly and the annual data with appropriate adjustment for heteroscedasicity. It is claimed that the coefficients obtained using pooled data will have a lower variance as compared to the estimators with yearly or quarterly data.

To show the relative efficiency of the estimates obtained using pooled data the author has made use of one of the standard theorems available in the econometric literature. There is thus little room to caste doubt on his claim. The paper, therefore, can be regarded as a theoretical contribution in the area.

A general problem faced by many researchers in at least developing countries is not that both quarterly as well as annual data are available for the same variables but that for some variables quarterly data are available while for others the annual data are available. In these circumstances the techniques suggested by the author is hardly of any use. I wonder if we can somehow modify the suggested technique to handle the above-mentioned problem which is more frequently faced by the researchers.

Nadeem A. Burney

Pakistan Institute of
Development Economics,
Islamabad