Econometrica, Vol. 39, No. 2 (March, 1971)

THE USE OF VARIANCE COMPONENTS MODELS IN POOLING CROSS SECTION AND TIME SERIES DATA

BY G. S. MADDALA[1]

The paper argues that variance components models are very useful in pooling cross section and time series data because they enable us to extract some information about the regression parameters from the between group and between time-period variation—a source that is often completely eliminated in the commonly used dummy variable techniques. The paper studies the applicability and usefulness of the maximum likelihood method and analysis of covariance techniques in the analysis of this type of model, particularly when one of the covariates used is a lagged dependent variable.

INTRODUCTION

THE USE OF analysis of covariance techniques in the problem of pooling cross section and time series data has now become a common practice in econometric work. Suppose we have data on N firms over T periods of time. The model usually used in pooling procedures is

$$y_{ij} = \alpha_i + \tau_j + \sum_{r=1}^{k} \beta_r x_{rij} + u_{ij} \qquad (i = 1, 2, \ldots N; j = 1, 2, \ldots T),$$

where α_i are the firm "dummies," τ_j are the time "dummies," and x_r are the "covariates." One common argument that is made against the use of the dummy variable technique is that it eliminates a major portion of the variation among both the explained and explanatory variables if the between firm and between time-period variation is large. In some cases there is also a loss in a substantial number of degrees of freedom. Added to these is the basic problem that rarely is it possible to give a meaningful interpretation to the dummy variables.

As a general approach to these problems, economists have now shifted their attention to models which treat the α_i (and τ_j) as random—in which case we estimate, instead of the $N\alpha$'s, only two parameters, the mean and variance of the distribution of the α's (and similarly for the time effects).[2] As far as the estimation of the slope parameters β's is concerned, this procedure amounts to extracting some information on the β's from the between firm and between time-period variation of the dependent and independent variables. We can also rationalize this procedure of treating the α_i and τ_j as random by arguing that the dummy variables do in effect represent some ignorance—just like the residuals u_{ij}. There is no reason to believe that this type of ignorance, which we might call "specific ignorance," should be treated differently than the "general ignorance" u_{ij}.

An earlier paper by Wallace and Hussain [**4**] analyzes this type of model and compares it with ordinary least squares and least squares with dummy variables.

[1] This research has been financed by the National Science Foundation under grant GS-1884. I wish to thank Marc Nerlove and Zvi Griliches for helpful comments. Responsibility for any errors is my own.

[2] For a study of this sort, see Balestra and Nerlove [**1**].

Wallace and Hussain do not, however, consider the case of lagged dependent variables—a case that Nerlove is worried about [**1, 2, 3**]. The present paper investigates some aspects of the analysis of variance components models that arise from the use of likelihood methods and the presence of lagged dependent variables as covariates. In particular the applicability and inapplicability of the usual analysis of covariance techniques will also be discussed.

The plan of the paper is as follows: Section 1 presents the model and the properties of the generalized least squares (GLS) estimates for a model with only firm effects. In Section 2 we study the behavior of the likelihood function and in Section 3 the applicability of the usual analysis of covariance techniques in the presence and absence of lagged dependent variables. In Section 4 an example is given to illustrate the nature of the biases discussed in Section 3. Section 5 presents an extension to the case of random firm and time effects. Section 6 presents an extension of the techniques to simultaneous equations methods. The final section presents the conclusions of the paper.

1. THE MODEL AND THE GLS ESTIMATE

Suppose we have observations on N individuals over T periods of time. The model we consider is $y = X\beta + u$ where y is an $NT \times 1$ vector, X is an $NT \times k$ matrix on k variables which may be exogenous or lagged dependent, β is a $k \times 1$ vector, u an $NT \times 1$ vector.

We can write the residuals as

$$u_{ij} = \mu_i + \tau_j + v_{ij},$$

where μ_i are the firm effects which are $IN\ (0, \sigma_\mu^2)$, τ_j are the time effects which are $IN\ (0, \sigma_\tau^2)$ and v_{ij} are $IN\ (0, \sigma_v^2)$. We assume the μ_i, τ_j, and v_{ij} to be independent. For the purpose of using least squares methods and analysis of covariance techniques we do not need the assumption of normality; nevertheless we will make it since we need it for ML methods. Also, for ease of exposition, we will omit the time effects. If this is done, we have

$$E(uu') = \Omega = \sigma^2 \begin{bmatrix} A & 0 & \dots & 0 \\ 0 & A & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A \end{bmatrix}$$

where A is the $T \times T$ matrix

$$\begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

and $\sigma^2 = \sigma_\mu^2 + \sigma_v^2$, $\rho = \sigma_\mu^2/\sigma^2$. It is evident that $A = (1 - \rho)I + \rho ee'$ and hence

$$A^{-1} = \lambda_1 ee' + \lambda_2 I \tag{1.1}$$

where $\lambda_1 = -\rho/(1 - \rho)(1 - \rho + T\rho)$, $\lambda_2 = 1/(1 - \rho)$, and e is a $T \times 1$ vector with all elements unity. We can also write our model as

$$y_i = X_i\beta + u_i \qquad (i = 1, 2 \ldots N).$$

Given the assumptions we make, the GLS estimate of β, if ρ is known, is

$$\hat{\beta} = \left[\sum_{i=1}^{N} X_i' A^{-1} X_i\right]^{-1}\left[\sum_{i=1}^{N} X_1' A^{-1} y_i\right].$$

Noting the expression for A^{-1} given in (1.1), we see that

$$\sum_{i=1}^{N} X_i' A^{-1} X_i = \lambda_1 \sum_{i=1}^{N} X_i' ee' X_i + \lambda_2 \sum_{i=1}^{N} X_i' X_i. \tag{1.2}$$

Define

$$T_{xx} = \sum_{i=1}^{N} X_i' X_i,$$

$$B_{xx} = \frac{1}{T}\sum_{i=1}^{N} (X_i' ee' X_i),$$

$$W_{xx} = T_{xx} - B_{xx},$$

with similar expressions for T_{xy}, B_{xy}, and W_{xy}. These are familiar expressions in analysis of variance. The matrix B_{xx} contains the sums of squares and sums of products between groups, W_{xx} is the corresponding matrix within groups, and T_{xx} is the corresponding matrix for total variation. Now (1.2) can be conveniently written as

$$\hat{\beta} = [W_{xx} + \theta B_{xx}]^{-1}[W_{xy} + \theta B_{xy}], \tag{1.3}$$

where $\theta = 1 + (\lambda_1 T/\lambda_2) = (1 - \rho)/(1 - \rho + \rho T) = \sigma_v^2/(\sigma_v^2 + T\sigma_\mu^2)$.

We see immediately that for fixed N, as $T \to \infty$, we have $\theta \to 0$ and if $((1/NT)W_{xx})$ and $((1/NT)W_{xy})$ have finite (non-null) probability limits, then plim $\hat{\beta}$ = plim $((1/NT)W_{xx})^{-1}((1/NT)W_{xy})$. But $((1/NT)W_{xx})^{-1}((1/NT)W_{xy})$ is the least squares estimate with dummy variables (hereafter to be denoted by LSDV). Also for fixed N and T, if $\rho \to 0$, then $\theta \to 1$, and $\hat{\beta} \to T_{xx}^{-1} T_{xy}$, which is the ordinary least squares (OLS) estimate. If $\rho \to 1$, $\theta \to 0$, and $\hat{\beta} \to W_{xx}^{-1} W_{xy}$, which is the LSDV estimate.

In essence θ measures the weight given to the between group variation. In the LSDV procedure, this source of variation is completely ignored. The OLS procedure corresponds to $\theta = 1$. Table I illustrates how θ varies with ρ for $T = 10$ and $T = 20$. As the table indicates, in the lower range of ρ, errors in the estimation of ρ will produce large changes in θ and hence result in substantial errors in the estimation of β, if the between group variation is large. The between group variation is large, however, only when ρ is large, and in this range errors in the estimation of ρ do not produce any substantial changes in θ.

TABLE I

$T = 10$		$T = 20$	
ρ	θ	ρ	θ
.05	.655	.05	.487
.1	.473	.1	.310
.2	.286	.2	.167
.3	.190	.3	.105
.4	.130	.4	.070
.5	.091	.5	.048
.6	.063	.6	.032
.7	.041	.7	.021
.8	.024	.8	.012
.9	.011	.9	.005

Formula (1.3) in essence combines the between group regression estimate and the within group regression estimate of β by weighting them in inverse proportion to their respective variances. In case there are only exogenous regressors both these estimates are evidently unbiased and the estimate of β given by (1.3) is the best unbiased linear estimate (if ρ is known). It is thus evidently more efficient than the LSDV estimate or the OLS estimate. In case there are lagged dependent variables, neither of these estimates is unbaised. It will be shown in Section 3 that in general the biases of these two estimates run in different directions and the process of combining the between group and within group regression estimates poses more problems.

A strong intuitive argument can be made for the pooling procedure suggested by (1.3). The usual procedures of OLS and LSDV are somewhat all or nothing ways of utilizing the between group variation. In the LSDV method, the between group variation is completely ignored. In OLS, the between group and within group variation is just added up. Usually, in pooling cross section and time series data, a test of significance is applied to test whether the constant terms are significantly different from each other. If the null hypothesis is rejected, one resorts to the LSDV method. If the null hypothesis is not rejected, one uses OLS. The GLS procedure implied in (1.3) is a compromise solution to this all or nothing way of utilizing the between group variation. Thus the procedure of treating the individual constant terms as random is a solution intermediate to treating them all as different and treating them all as equal. A similar argument can be made even for those procedures that treat the slope coefficients as random.

2. ML ESTIMATION

If ρ is known, then the estimation of β is straightforward. The estimate is given by (1.3), which is also the ML estimate. In case there are lagged dependent variables among the regressors, we have to make some assumptions about the initial values of the y's, but this does not introduce any essential complications into this estimation problem.

If ρ is not known, two possibilities suggest themselves: (i) use the ML method; (ii) use analysis of covariance techniques to get unbiased estimates of σ_μ^2 and σ_v^2 and use these in a two-step GLS procedure. If lagged dependent variables are present, procedure (ii) is ruled out because the analysis of covariance does not give unbiased estimates of σ_μ^2 and σ_v^2. In fact the between group mean square is a seriously biased estimate of the between group variance $\sigma_v^2 + T\sigma_\mu^2$. But some modifications of this procedure will be considered later. The ML method can be applied even if lagged dependent variables are present. However, a study of the behavior of the likelihood function in this model would be very fruitful.

After differentiating the likelihood function with respect to β, we obtain the ML estimate of β as the expression given in (1.3), and substituting this in the likelihood function we get

$$-2\log L = \text{const} + NT\log\sigma_v^2 + N\log\left[1 + \frac{T\rho}{1-\rho}\right] + \frac{1}{\sigma_v^2}[W_{yy} + \theta B_{yy} - (W_{yx} + \theta B_{yx})(W_{xx} + \theta B_{xx})^{-1}(W_{xy} + \theta B_{xy})] \tag{2.1}$$

where $\theta = (1-\rho)/(1-\rho+T\rho)$ and the other expressions are as defined earlier. Differentiating this with respect to σ_v^2 yields

$$NT\hat{\sigma}_v^2 = [W_{yy} + \theta B_{yy} - (W_{yx} + \theta B_{yx})(W_{xx} + \theta B_{xx})^{-1})(W_{xy} + \theta B_{xy})]. \tag{2.2}$$

Hence, by substitution, the concentrated likelihood function in terms of ρ only is given as

$$-2\log L = \text{const} + NT\log\hat{\sigma}_v^2 + N\log\left[1 + \frac{T\rho}{1-\rho}\right]. \tag{2.3}$$

This expression is now a function of ρ only. Differentiating this with respect to ρ gives us the ML estimate of ρ. However, the expression is not easy to manipulate. The behavior of expression (2.3) as ρ ranges from 0 to 1 can be more conveniently studied by computing the derivatives of (2.3) with respect to θ given in (2.1). Note that as ρ increases from 0 to 1, θ decreases from 1 to 0.

Now $\log[1 + (T\rho/(1-\rho))] = -\log\theta$, and $\log\theta$ is a steadily increasing function of θ. As for the behavior of $\hat{\sigma}_v^2$, we have

$$\begin{aligned}\frac{\partial(NT\hat{\sigma}_v^2)}{\partial\theta} &= B_{yy} - [B_{yx}(W_{xx} + \theta B_{xx})^{-1}(W_{xy} + \theta B_{xy}) - (W_{yx} + \theta B_{yx}) \\ &\quad \times (W_{xx} + \theta B_{xx})^{-1}B_{xx}(W_{xx} + \theta B_{xx})^{-1}(W_{xy} + \theta B_{xy}) \\ &\quad + (W_{yx} + \theta B_{yx})(W_{xx} + \theta B_{xx})^{-1}B_{xy}] \\ &= [1, \lambda']\begin{bmatrix} B_{yy} & B_{yx} \\ B_{xy} & B_{xx}\end{bmatrix}\begin{bmatrix}1 \\ \lambda\end{bmatrix}\end{aligned}$$

where $\lambda = (W_{xx} + \theta B_{xx})^{-1}(W_{xv} + \theta B_{xv})$. Since the B matrix is positive definite, this expression is greater than zero. Hence we have $\partial\hat{\sigma}_v^2/\partial\theta > 0$ or $\partial\log\hat{\sigma}_v^2/\partial\theta > 0$.

Also noting that $(\partial\lambda/\partial\theta) = -[W_{xx} + \theta B_{xx}]^{-1}[B_{xx}\lambda - B_{xy}]$, we have

$$\partial^2(NT\hat{\sigma}_v^2)/\partial\theta^2 = 2[\lambda' B_{xx}(\partial\lambda/\partial\theta) - B_{yx}(\partial\lambda/\partial\theta)] = -2(\lambda' B_{xx} - B_{yx}) \times [W_{xx} + \theta B_{xx}]^{-1}(B_{xx}\lambda - B_{xy})$$

which is less than zero for $\theta > 0$.

Thus the first derivative of $NT \log \bar{\sigma}_v^2$ with respect to θ is greater than zero and the second derivative with respect to θ is less than zero. In Figure 1, curve I shows the behavior of $NT \log \hat{\sigma}_v^2$ (assuming $\hat{\sigma}_v^2 > 1$) and curve II shows the behavior of $\log \theta$. The position of curve II is fixed. The position of curve I will vary but in any case, for finite N and T, $NT \log \hat{\sigma}_v^2$ has to be finite through the range 0 to 1 (as in curve III).

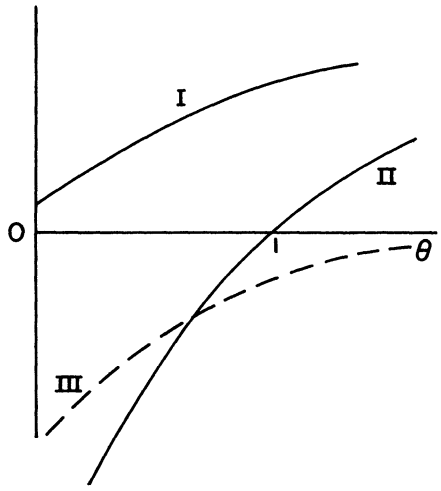

FIGURE 1

Three conclusions follow from this.

(i) The maximum of the likelihood function cannot occur at the boundary value $\theta = 0$ or $\rho = 1$.

(ii) Since the maxima of the likelihood function correspond to the points where the distance between curves I and II is minimum, the number of maxima will depend on the relative curvature of the two curves. Since both the functions are increasing at a decreasing rate, however, there can be at most two maxima for the likelihood function in the range $0 < \theta \leqslant 1$. (This implies that we have to guard against one local maximum.)

(iii) If we confine θ to the range $0 < \theta \leqslant 1$, a necessary and sufficient condition for the occurence of a boundary solution at $\theta = 1$ (i.e., $\rho = 0$) is

$$\frac{\partial}{\partial\theta}[T \log \hat{\sigma}_v^2 - \log\theta]_{\theta=1} < 0,$$

i.e.,

$$T_{yy} - \alpha' T_{xx}\alpha > T[B_{yy} - 2\alpha' B_{xy} + \alpha' B_{xx}\alpha] \tag{2.4}$$

where $\alpha = T_{xx}^{-1}T_{xy}$.

Note, however, that the ML estimate of ρ and the estimate obtained from the analysis of covariance are not the same. Hence there can be situations when the former method gives a boundary solution $\rho = 0$ whereas the latter method does not, and conversely.

For instance, consider the case $N = 25$ and $T = 10$. If $B_{yy} = 113$, $B_{xy} = 60$, $B_{xx} = 40$, $W_{yy} = 264$, $W_{xy} = 40$, and $W_{xx} = 40$, then the analysis of variance estimate of ρ is 0; but the likelihood function is increasing at $\rho = 0$ and hence the ML estimate of ρ is greater than zero. On the other hand, if $B_{yy} = 83$, $B_{xy} = 60$, $B_{xx} = 40$, $W_{yy} = 240$, $W_{xy} = 40$, and $W_{xx} = 40$, then the analysis of variance estimate of ρ is greater than zero since condition (2.4) is satisfied, we get a boundary solution at $\rho = 0$ by the ML method.

3. ANALYSIS OF COVARIANCE ESTIMATES

In the case where all the x's are exogenous, we can easily show that both the between group estimate $B_{xx}^{-1}B_{xy}$ and the within group estimate $W_{xx}^{-1}W_{xy}$ are unbiased estimates of β. One could perform an analysis of covariance as shown in Table II. If we denote the between group mean square by BMS and the within group mean square by WMS, then WMS gives an unbiased estimate of σ_v^2 and $(BMS - WMS)/T$ gives an unbaised estimate of σ_μ^2.

TABLE II

Source	Covariance matrix	$\hat{\beta}$	Residual sum of squares	Degrees of freedom	E(RSS/d.f.)
Between group	$B_{yy}B_{yx}B_{xx}$	$B_{xx}^{-1}B_{xy}$	$B_{yy} - B_{yx}B_{xx}^{-1}B_{xy}$	$N - 1 - k$	$\sigma_v^2 + T\sigma_\mu^2$
Within group	$W_{yy}W_{yx}W_{xx}$	$W_{xx}^{-1}W_{xy}$	$W_{yy} - W_{yx}W_{xx}^{-1}W_{xy}$	$N(T-1) - k$	σ_v^2
Total	$T_{yy}T_{yx}T_{xx}$	$T_{xx}^{-1}T_{xy}$	$T_{yy} - T_{yx}T_{xx}^{-1}T_{xy}$	$NT - 1 - k$	

There is no guarantee that the estimate of σ_μ^2 will be positive. In fact this is the familiar problem of negative variance components. One suggestion, common in analysis of variance literature, is to put $\hat{\sigma}_\mu^2 = 0$ if $BMS < WMS$, which amounts to using the OLS estimate of β in this case. The other possibility is to say that the WMS is high because it "captures" variation due to other omitted effects (e.g., the time effects, if they are not already included), in which case the proper solution is to go back and examine the model and correct it for the omitted variables.

In any case, as was pointed out earlier, the $\hat{\beta}$ obtained by pooling the between group estimate and the within group estimate using the variables estimated from the analysis of covariance does not give the ML estimate. The estimate we will be using is

$$\hat{\beta} = [W_{xx} + \hat{\theta}B_{xx}]^{-1}[W_{xy} + \hat{\theta}B_{xy}] \tag{3.1}$$

where $\hat{\theta} = \hat{\sigma}_v^2/(T\hat{\sigma}_\mu^2 + \hat{\sigma}_v^2)$. We can write $\hat{\beta} = \beta + [W_{xx} + \hat{\theta}B_{xx}]^{-1}[W_{xu} + \hat{\theta}B_{xu}]$. Since the estimates of σ_v^2 and $(T\sigma_\mu^2 + \sigma_v^2)$ are obtained from the least squares residuals from the within group and between group regression, they can be easily shown to be independent of W_{xu} and B_{xu} and hence it will follow that $E(\hat{\beta}) = \beta$. Thus the GLS estimate is unbiased even if the variance components (and hence the covariance matrix of the residuals) are estimated.

In the presence of lagged dependent variables (the case Nerlove was concerned about), neither the between group regression nor the within group regression estimates are unbiased. It is first important to see whether we can say anything about the direction of the biases. It is easier to determine the direction of the biases for the between group regression estimates.

Consider the model:

$$y_{it} = \alpha y_{i,t-1} + \beta x_{it} + u_{it}.$$

The between group regression estimates of α and β are computed from

$$y_i = \alpha Z_i + \beta x_i + u_i \qquad (i = 1, 2, \ldots N),$$

where $y_i = \Sigma_t y_{it}$, $x_i = \Sigma_t x_{it}$, $Z_i = \Sigma_t y_{i,t-1}$. Noting that y_i and Z_i have $(T-1)$ observations in common, we would expect them to be highly correlated. Hence if M denotes the matrix of covariances of y_i, Z_i, x_i,

$$\begin{bmatrix} M_{yy} & M_{yZ} & M_{yx} \\ & M_{ZZ} & M_{Zx} \\ & & M_{xx} \end{bmatrix},$$

then unless T is very small, we would expect $M_{yy} \simeq M_{yZ} \simeq M_{ZZ}$ and $M_{yx} \simeq M_{Zx}$ (where $\simeq$ denotes "approximately equal to"). Since

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{1}{\Delta}\begin{bmatrix} M_{xx} & + M_{Zx} \\ + M_{Zx} & M_{ZZ} \end{bmatrix}\begin{bmatrix} M_{yx} \\ M_{yZ} \end{bmatrix},$$

where $\Delta = M_{ZZ}M_{xx} - M_{Zx}^2$, we would expect the estimate $\hat{\alpha}$ to be close to 1 and the estimate $\hat{\beta}$ to be close to 0. We would also expect the between group mean square to be seriously biased downwards because of the high correlation between y_i and Z_i.

Hence, if we start with the true values $0 < \alpha < 1$ and $\beta > 0$, we will find the between group regression estimates of α upward biased and those of β downward biased. Nerlove, in his Monte Carlo studies [**2, 3**], does not report the estimates of the between group regression. But the example in the next section illustrates these conclusions.

The direction of biases for the within group regression, viz., the LSDV, is not so easy to analyze. However, we can say something if we are able to prove that the elements of the vector $\beta(\theta) = [W_{xx} + \theta B_{xx}]^{-1}[W_{xy} + \theta B_{xy}]$ are monotonically increasing or decreasing functions of θ. Noting that $\theta = (1 - \rho)/(1 - \rho + \rho T)$ we see that for $\theta = 0$, we have the LSDV estimate; for some θ in the range $0 < \theta < 1$,

we have the GLS estimate; for $\theta = 1$, we have the OLS estimate; and for θ very large, we have the between group regression estimates. Also noting that the between group regression coefficient of α is close to 1, and the coefficient of β is close to 0, if $\hat{\alpha}(\theta)$ is a monotonically increasing function of θ and $\hat{\beta}(\theta)$ is a monotonically decreasing function of θ, then we can show that the OLS estimate of α will be upward biased and of β downward biased; and the LSDV estimate of α will be downward biased and of β upward biased, which is what Nerlove finds in his Monte Carlo study [3].

It has not been possible to establish any general conclusions of this sort because the condition involves some complicated expressions involving the within group and between group covariances. In the case where there is only one covariate, it is very easy to show that $\hat{\beta}(\theta)$ is a monotonically increasing or decreasing function of θ. In the general case too, it appears that a similar result holds good in a large number of situations, at least within the range $0 < \theta \leqslant 1$. The example given in the next section illustrates these conclusions about the biases in the within group and the between group regressions, and the monotonic behavior of $\hat{\alpha}(\theta)$ and $\hat{\beta}(\theta)$.

Since the between group regression and the within group regression are both biased (and the biases run in opposite directions), we should be able to do better by taking a linear combination of these two regressions. Now it is also clear that if the between group mean square is biased downwards, giving these two regression estimates weights inversely proportional to their variances will give unduly heavy weight to the between group regression. Hence we cannot rely on any estimates obtained from the analysis of covariance similar to that mentioned in Table II.

In the case of ML too, noting that the between group covariance matrix will be close to singularity in the presence of lagged dependent variables, the condition (2.4) will also be satisfied more often. Hence the ML method will also give boundary solutions more often in this case than in the case of purely exogenous variables.

In view of these results, Nerlove has suggested an alternative procedure. His suggestion is to use the within group mean square as an estimate of σ_v^2 and to estimate σ_μ^2 as the variance of the estimated dummies in the within group regression. This estimate can be written as

$$\hat{\sigma}_\mu^2 = \frac{1}{N}\sum_{i=1}^{N}\left[y_{i\cdot} - y_{\cdot\cdot} - \sum_{r=1}^{k}\hat{\beta}_r(x_{ri} - x_{r\cdot\cdot})\right]^2$$

where

$$y_{i\cdot} = \frac{1}{T}\sum_t y_{it}, \qquad y_{\cdot\cdot} = \frac{1}{N}\sum_i y_{i\cdot},$$

with similar expressions holding for the k covariates $x_{r\cdot}$; $\hat{\beta}_r$ are the estimates of β_r obtained from the within group regression.

In the case where the x's are all exogenous, the expected value of this estimate is $\sigma_\mu^2 + (\sigma_v^2/T)$ which is slightly upward biased. Though this estimate is upward biased, it has the advantage of being always positive.

In the case where there are lagged dependent variables, the bias in this estimate is harder to evaluate, particularly because the $\hat{\beta}_r$ obtained from the within group

regression are themselves biased. The indication from Nerlove's Monte Carlo study [3] is that it is strongly biased upwards. As for $\hat{\sigma}_v^2$ too, it is no longer an unbiased estimate of σ_v^2 in the presence of lagged dependent variables.

4. AN ILLUSTRATIVE EXAMPLE

The following artificial example illustrates the results presented in the previous sections. In view of the fact that Nerlove has done extensive Monte Carlo studies [2, 3], we have investigated only one sample. The main purpose of this example is to investigate certain aspects of the model that were not investigated by Nerlove—in particular the results of the between group regressions, and the monotonic behavior of $\hat{\alpha}(\theta)$ and $\hat{\beta}(\theta)$, and the possibility of multiple maxima for the likelihood function.

Data were generated on the following models:

$$y_{it} = 1.0 + \alpha y_{i,t-1} + u_{it} \quad \text{(Model 1)},$$

$$y_{it} = \beta x_{it} + u_{it} \quad \text{(Model 2)},$$

$$y_{it} = \alpha y_{i,t-1} + \beta x_{it} + u_{it} \quad \text{(Model 3)},$$

$$(i = 1, 2, \ldots N;\ t = 1, 2, \ldots T).$$

Model 1 has a lagged dependent variable, Model 2 has an exogenous variable, and Model 3 has both. The u_{it} are assumed to have a normal distribution with covariance matrix Ω given in Section 1. The parameters chosen were: $\alpha = 0.7$, $\beta = 0.5$, $\sigma^2 = 1.0$, $\rho = 0.4$, $N = 25$, and $T = 10$. The exogenous variables were picked up from Nerlove's Monte Carlo study [3]. The same u_{it} were used for all models. The initial values y_{i1} were taken as $y_{i1} = u_{i1}/\sqrt{1-\alpha^2}$ for Model 1 and $y_{i1} = \beta x_{i1} + u_{i1}/\sqrt{1-\alpha^2}$ for Model 3. In all, twenty values were generated for each i, but the first ten were discarded so that $T = 10$. The results of the within group and between group regressions were as follows.

For Model 1,

within group: $\hat{\alpha} = .4747$, $R^2 = .48$, M. Sq. $= .1211$;
between group: $\hat{\alpha} = .9888$, $R^2 = .93$, M. Sq. $= 2.7433$.

For Model 2,

within group: $\hat{\beta} = .3422$, $R^2 = .10$, M. Sq. $= .48$;
between group: $\hat{\beta} = .4710$, $R^2 = .85$, M. Sq. $= 3.6393$.

For Model 3,

within group: $\hat{\alpha} = .3178$, $\hat{\beta} = 1.0535$, $R^2 = .80$, M. Sq. $= .26$;
between group: $\hat{\alpha} = 1.000$, $\hat{\beta} = .1032$, $R^2 = .9994$, M. Sq. $= .12$.

Note that the between group estimate of α is biased towards 1 as expected. Further, condition (2.4), for the occurrence of a boundary solution by the ML method is satisfied only in Model 3. This is also the case where the within group

mean square is greater than the between group mean square and hence the analysis of covariance estimate of ρ is negative (though, as discussed in the text, there is no basis for the use of this method in the presence of lagged dependent variables). Both the estimates $\hat{\alpha}$ and $\hat{\beta}$ are functions of θ.

In this particular example

$$\frac{\partial \hat{\alpha}(\theta)}{\partial \theta} = 3302.89\,\theta^2 + 334.76\,\theta + 8.62$$

which is greater than 0 for $\theta > 0$. Also $\partial \hat{\beta}(\theta)/\partial\theta = -4376.88\,\theta^2 - 466.32\,\theta - 13.43$ which is less than 0 for $\theta > 0$. Hence, $\bar{\alpha}(\theta)$ is a monotonically increasing function of θ and $\hat{\beta}(\theta)$ is a monotonically decreasing function of θ.

Finally, to investigate whether the boundary solution at $\rho = 0$ in Model 3 gives a local maximum or a global maximum, the likelihood function was tabulated over the entire range $\rho = -.10$ to $\rho = .99$ at intervals of 0.01. It was found that the maximum at $\rho = 0$ was in fact a local maximum. The values of $\log L$ up to a constant are shown in Table III for selected values of ρ. There is a local maximum

TABLE III

$A = \log L$ + const. for Model 3[a]

ρ	A	ρ	A
−.10	−24.68	.50	−6.96
−.09	−15.67	.60	−3.75
−.08	−13.80	.70	−0.76
−.07	−13.34	.72	−0.24
−.06	−13.50	.74	0.24
−.05	−13.75	.76	0.67
−.04	−14.12	.78	1.05
−.03	−14.42	.80	1.33
−.02	−14.75	.82	1.52
−.01	−15.03	.84	1.58
.00	−15.27	.86	1.47
.05	−16.05	.88	1.12
.10	−16.12	.90	0.45
.15	−15.73	.92	−0.70
.20	−15.01	.94	−2.63
.25	−14.02	.96	−6.01
.30	−12.85	.98	−12.95

[a] There are two maxima, a local maximum at $\rho = -.07$ and a global maximum at $\rho = .84$.

at $\rho = -0.7$ and a global maximum at $\rho = .84$. In the case of Models 1 and 2 there was only one maximum for the likelihood function. The relative likelihoods of ρ for the three models are plotted in Figure 2. They do confirm the large bias in the ML estimate of ρ that Nerlove talks about, though the direction of the bias is towards the LSDV estimate rather than towards $\rho = 0$. This bias is large for the models with lagged dependent variables but not for the model with only exogenous

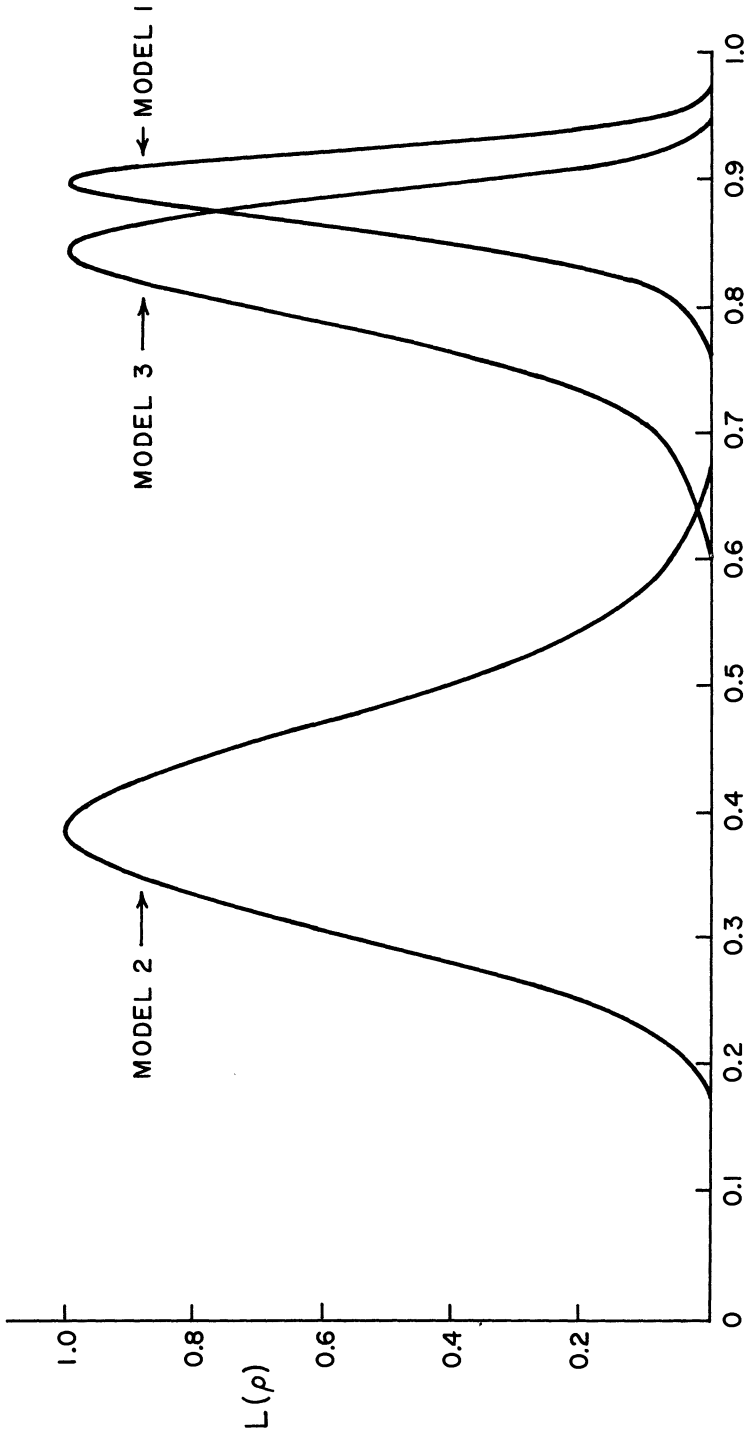
MODEL 1
MODEL 3
MODEL 2
L(ρ)
ρ
1.0
0.8
0.6
0.4
0.2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

FIGURE 2

variables. Further, the likelihood functions are not flat around the maximum, thus indicating that the problem is not one of likelihood functions spread over a wide range of values of ρ and of our picking a maximum point on the likelihood function that cannot be adequately distinguished from other values. Finally, in Model 3 there was a local maximum at $\rho = -0.7$ but the relative likelihood at this point (relative to the value of the likelihood function at the global maximum $\rho = .84$) was negligibly small. It is important to guard against such local maxima in finding the ML estimate.

5. ANALYSIS OF THE MODEL WITH RANDOM TIME EFFECTS

The analysis contained in the previous sections can be very easily extended to the case where there are random time effects in addition to random firm effects. The rationalization for treating the time dummies as random is precisely the same as that for treating the firm dummies as random.

The model now is $y = X\beta + u$ where u is an NT component vector, the i,jth element u_{ij} being equal to $\mu_i + \tau_j + v_{ij}$ where μ_i are the firm effects and τ_j are the time effects. We shall assume that μ_i are $IN(0, \sigma_\mu^2)$, τ_j are $IN(0, \sigma_\tau^2)$, v_{ij} are $IN\ (0, \sigma_v^2)$, and that these are mutually independent. We shall also assume that all variables are measured as deviations from their respective means.

We can, as before, decompose the variances and covariances (let us call these T_{xx}, T_{xy}, and T_{yy}) into three parts: (i) between firms (let us call these B_{xx}, B_{xy}, B_{yy}); (ii) between time periods (let us call these C_{xx}, C_{xy}, C_{yy}); and (iii) the residual (let us call these W_{xx}, W_{xy}, W_{yy}). Define $\theta_1 = \sigma_v^2/(\sigma_v^2 + T\sigma_\mu^2)$ and $\theta_2 = \sigma_v^2/(\sigma_v^2 + N\sigma_\tau^2)$. Then the GLS estimate of β can be easily shown to be equal to

$$\hat{\beta} = [W_{xx} + \theta_1 B_{xx} + \theta_2 C_{xx}]^{-1}[W_{xy} + \theta_1 B_{xy} + \theta_2 C_{xy}]. \tag{5.1}$$

This is a generalization of formula (1.3). As $N \to \infty$ and $T \to \infty$, $\hat{\beta} \to W_{xx}^{-1} W_{xy}$, which is the LSDV estimate. For $\theta_1 = 1$ and $\theta_2 = 1$, we get the OLS estimate.

What formula (5.1) does is to combine the regression estimates obtained from the between firm variation, between time-period variation and the residual variation, weighting them in inverse proportion to their variances. In case there are only exogenous variables present, all three of these estimates are unbiased and the estimate of β given by (5.1) is the best unbiased linear estimate (if θ_1 and θ_2 are known). It is thus evidently more efficient than the LSDV or OLS estimates. In case there are lagged dependent variables present, however, none of these three estimates is unbiased. As before we can argue that the between firm and between time-period regression estimates are biased, the coefficient of the lagged dependent variable being biased towards 1 and the coefficient of the exogenous variable being biased towards 0. Again, the intuitive argument in favor of the pooling procedure suggested by (5.1) is that the usual procedures of OLS and LSDV are all or nothing ways of utilizing the between time-period variation and the GLS procedure implied in (5.1) is a compromise solution.

The analysis of the behavior of the likelihood function contained in Section 2 can be extended to this case. But no simple conclusions are possible because of

the presence of some interaction terms. First, we note that the determinant of the covariance matrix Ω is given by

$$\log|\Omega| = (N-1)(T-1)\log\sigma_v^2 + (N-1)\log(\sigma_v^2 + T\sigma_\mu^2) + (T-1)\log(\sigma_v^2 + N\sigma_\tau^2) + \log(\sigma_v^2 + T\sigma_\mu^2 + N\sigma_\tau^2)$$

which can be written in terms of θ_1, θ_2, and σ_v^2. After the simplifications used in Section 2 we can write the concentrated likelihood function as

$$-2\log(\theta_1,\theta_2) = \text{const.} + NT\log\hat{\sigma}_v^2 - N\log\theta_1 - T\log\theta_2 + \log(\theta_1 + \theta_2 - \theta_1\theta_2) \tag{5.2}$$

where

$$NT\hat{\sigma}_v^2 = [W_{yy} + \theta_1 B_{yy} + \theta_2 C_{yy} - (W_{yx} + \theta_1 B_{yx} + \theta_2 C_{yx}) \times (W_{xx} + \theta_1 B_{xx} + \theta_2 C_{xx})^{-1}(W_{xy} + \theta_1 B_{xy} + \theta_2 C_{xy})]. \tag{5.3}$$

Now $\log\theta_1$ is a steadily increasing function of θ_1 and $\log\theta_2$ is a steadily increasing function of θ_2. Also, as before, we can easily show that

$$\frac{\partial(NT\hat{\sigma}_v^2)}{\partial\theta_1} = [1, \lambda']\begin{bmatrix} B_{yy} & B_{yx} \\ B_{xy} & B_{xx} \end{bmatrix}\begin{bmatrix} 1 \\ \lambda \end{bmatrix}$$

and

$$\frac{\partial(NT\hat{\sigma}_v^2)}{\partial\theta_2} = [1, \lambda']\begin{bmatrix} C_{yy} & C_{yx} \\ C_{xy} & C_{xx} \end{bmatrix}\begin{bmatrix} 1 \\ \lambda \end{bmatrix}$$

where $\lambda = [W_{xx} + \theta_1 B_{xx} + \theta_2 C_{xx}]^{-1}[W_{xy} + \theta_1 B_{xy} + \theta_2 C_{xy}]$, and since the B and C matrices are positive definite, these expressions are greater than zero. Also

$$\frac{\partial^2(NT\hat{\sigma}_v^2)}{\partial\theta_1^2} = -2(\lambda' B_{xx} - B_{yx})[W_{xx} + \theta_1 B_{xx} + \theta_2 C_{xx}]^{-1}(B_{xx}\lambda - B_{xy})$$

and

$$\frac{\partial^2(NT\hat{\sigma}_v^2)}{\partial\theta_2^2} = -2(\lambda' C_{xx} - C_{yx})[W_{xx} + \theta_1 B_{xx} + \theta_2 C_{xx}]^{-1}(C_{xx}\lambda - C_{xy}),$$

which are less than zero for $\theta_1 > 0$ and $\theta_2 > 0$. Thus, the first derivative of $NT\log\hat{\sigma}_v^2$ with respect to θ_1 is greater than zero and the second derivative is less than zero, and similarly for the derivatives with respect to θ_2. But because of the presence of the interaction term $\log(\theta_1 + \theta_2 - \theta_1\theta_2)$, no simple conclusions such as those in Section 2 can be deduced. If we assume that this factor is dominated by the other factors in (5.3) so that it can be ignored, then we can show that a boundary solution will occur (i) at $\theta_1 = 1$ if

$$W_{yy} + B_{yy} + \theta_2 C_{yy} - \alpha_1'(W_{xx} + B_{xx} + \theta_2 C_{xx})\alpha_1 > T(B_{yy} - 2\alpha_1' B_{xy} + \alpha_1' B_{xx}\alpha_1),$$

or (ii) at $\theta_2 = 1$ if

$$W_{yy} + \theta_1 B_{yy} + C_{yy} - \alpha_2'(W_{xx} + \theta_1 B_{xx} + C_{xx})\alpha_2 > N(C_{yy} - 2\alpha_2' C_{xy} + \alpha_2' C_{xx}\alpha_2)$$

where

$$\alpha_1 = (W_{xx} + B_{xx} + \theta_2 C_{xx})^{-1}(W_{xy} + B_{xy} + \theta_2 C_{xy})$$

and

$$\alpha_2 = (W_{xx} + \theta_1 B_{xx} + C_{xx})^{-1}(W_{xy} + \theta_1 B_{xy} + C_{xy}).$$

In any case these conclusions are not very useful because there is no way of deciding a priori whether or not these boundary solutions correspond to global maxima of the likelihood function.

6. SIMULTANEOUS EQUATIONS MODELS ESTIMATED ON A TIME SERIES OF CROSS SECTIONS

The results in the previous sections can be easily extended to simultaneous equations models. Since the algebraic manipulations are similar, only the final results will be stated here.

If one is interested in estimating only the reduced form equations, since the residuals of the reduced form equations have the same structure as the residuals of the structural equations, we obtain three estimates for the reduced form parameters: one from the between firm variation, one from the between time-period variation, and one from the residual variation. Also by virtue of the well known result that simultaneous estimation of the system of unrestricted reduced form equations is equivalent to estimating each equation separately, we can use the variance component technique separately for each equation. If there are no lagged endogenous variables in the model, then one can use the analysis of variance described in Section 3 and get estimates of the variance components. The subsequent estimates of the reduced form parameters obtained on the basis of these estimated variance components are still unbiased. If there are lagged endogenous variables in the system, then, as mentioned earlier, none of the estimates (from the between firm, between time-period, and residual variation) is unbiased. The problem of optimal estimation of variance components has been discussed earlier and nothing needs to be added to the earlier analysis.

Things get complicated when it comes to structural estimation. Suppose we are interested in estimating the first structural equation by two-stage least squares (2SLS). Again, a decomposition of the error into three independent components (between firm, between time period, and residual) leads us to three independent estimates for the parameters concerned. One can decompose the variance-covariance matrix of the endogenous and exogenous variables into three components: (i) between firms—say the B matrix; (ii) between time periods—say the C matrix; and (iii) residual—say the W matrix. Let $T = B + C + W$. Then one can

compute 2SLS estimates from each of these covariance matrices. The 2SLS estimates obtained from the T matrix are the estimates from the pooled sample. The 2SLS estimates obtained from the W matrix are the estimates obtained if the data are pooled but firm and time-period dummies are introduced. If only firm dummies are used, then we use the $(W + C)$ matrix. The efficient variance component estimates are obtained by weighting the independent estimates from B, C, and W in inverse proportion to their variances.

The problem we run into, however, is that the covariance matrices of these three estimates (in addition to involving the unknown variance components) are the asymptotic covariance matrices, and if we resort to asymptotic arguments, we again encounter the old problem that the variance component estimator and the usual estimator with dummy variables are equivalent. In any case if one is faced with the problem of estimating a simultaneous equations model on the basis of data consisting of a time series of cross sections, it is advisable to compute in practice the 2SLS estimates (or any other estimates being used) from each of the above mentioned sources of variation in addition to the total. Often, it might happen that there would not be enough degrees of freedom available in the B matrix or the C matrix. If this is so, this matrix should be pooled with the W matrix.

Alternatively one could obtain the 2SLS estimates from the covariance matrix $(W + \theta_1 B + \theta_2 C)$ where θ_1 and θ_2 lie between 0 and 1. One could compute these estimates for different values of θ_1 and θ_2 (say at intervals of 0.1) and choose that set of estimates for which the generalized variance of the estimated covariance matrix is minimum. These techniques will be illustrated with an example in a subsequent paper.

7. CONCLUSIONS

The paper argues that variance components models are very useful in pooling cross section and time series data because they enable us to extract some information about the regression parameters from the between group and between time-period variation—a source that is often completely eliminated in the commonly used dummy variable techniques. We can also rationalize this procedure of treating the firm effects and time effects as random by arguing that these effects too, like the over-all residual, measure our ignorance and there is no reason to treat one source of ignorance as random and the other as fixed. The paper studies the applicability and usefulness of the maximum likelihood method and analysis of covariance techniques in the analysis of this type of model—particularly when one of the covariates used is a lagged dependent variable. The paper first analyzes a model with only random firm effects and then extends the analysis to one with random firm and time effects. Since the conclusions are similar, we shall summarize the conclusions for a model with only firm effects. There are four conclusions.

(1) When we write the likelihood function in its concentrated form, it consists of two components, one a steadily increasing function of the parameter ρ, and the other a steadily decreasing function. The likelihood function cannot attain a maximum at the boundary value $\rho = 1$ (corresponding to LSDV). But it can

attain a boundary solution at $\rho = 0$ (corresponding to OLS) if the range of ρ is confined to $0 \leqslant \rho < 1$. The condition for the occurrence of such a boundary solution has been derived. It shows that the boundary solution can occur even when the covariates are exogenous, though the boundary solution could occur more often if the covariates contain lagged dependent variables since the between group covariance matrix may then be close to singularity. This boundary solution, however, can correspond to just a local maximum rather than a global maximum as illustrated by the example in Section 4.

(2) As for the usefulness of covariance techniques, when only exogenous variables are present, both the between group and within group regressions give unbiased estimates of the slope coefficients. In general we could combine these two estimates by weighting them in inverse proportion to their variances, as obtained from the usual analysis of covariance. This does not amount to using the ML method. Pooling on the basis of estimated variances and pooling with the use of the likelihood function are not the same.

(3) In the case where there are lagged dependent variables, neither the between group nor the within group regression gives unbiased estimates of the slope coefficients. The between group regresson estimates are badly biased—the coefficient of the lagged dependent variable is biased towards one and the coefficient of the exogenous variable towards zero. Also, the between group mean square is biased downwards. Hence analysis of covariance techniques cannot be relied on to give optimal estimates. The method of ML, too, has its drawbacks since it often gives boundary solutions. The solution offered by Nerlove does not get us into this problem, and as shown by his Monte Carlo studies, it gives better estimates than the method of ML or the LSDV method. However, as our analysis indicates, it is expected to be biased towards the LSDV method.

(4) Those working with problems of pooling cross section and time series data usually present either the OLS or the LSDV estimates. It is, however, advisable to present, in addition, the estimates obtained from the between group and between time-period variation.

Since the analysis in this paper can be easily extended to simultaneous equations models based on time series of cross sections, these conclusions hold good for such models too. One important assumption, however, that is needed for the validity of the analysis in this paper is that the "specific ignorance" be assumed to be independent of the regressors—an assumption that may not always be valid and that is not needed for the consistency of the least squares with dummy variable techniques.

University of Rochester

Manuscript received May, 1969; revision received October, 1969.

REFERENCES

[1] BALESTRA, P., AND M. NERLOVE: "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas," *Econometrica* (July, 1966), 585–612.

[2] NERLOVE, M.: "Experimental Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections," *Economic Studies Quarterly* (December, 1967), 42–74.
[3] ———: "Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections," *Econometrica*, this issue, pp. 359–382.
[4] WALLACE, T. D., AND A. HUSSAIN: "The Use of Error Components Models in Combining Cross Section with Time Series Data," *Econometrica* (January, 1969), 55–72.