Cross Sectional Time Series #1

Paul Johnson <pauljohn@ku.edu>

5th April 2004

Suppose your dependent variable is Normally distributed. What are you supposed to do?

1 The Longitudinal Data Problem

Longitudinal data sets include repeated observations on each of many units, such as nations, states, counties, or cities.

So we use 2 subscripts for data, the first refers to the unit –or "cluster"–the second to the time. A vector of coefficients is b and the dependent variable is y_{it} and the set of independent variables observed for each country and time is x_{it}. The model looks something like:

$$
\begin{array}{lcll}
y_{11} & & x_{11}b & + \quad e_{11} \\
y_{12} & & x_{12}b & + \quad e_{12} \\
y_{13} & & x_{13}b & + \quad e_{13} \\
\ldots & & & + \\
y_{1T} & & x_{1T}b & + \quad e_{1T} \\
y_{21} & = & x_{21}b & + \quad e_{21} \\
y_{22} & & x_{22}b & + \quad e_{22} \\
\ldots & & & + \\
y_{2T} & & x_{2T}b & + \quad e_{2T} \\
y_{31} & & x_{31}b & + \quad e_{31} \\
y_{32} & & x_{32}b & + \quad e_{32} \\
\ldots & & & + \\
y_{3T} & & x_{3T}b & + \quad e_{3T}
\end{array}
$$

What is the variance of the error term? And how do deviations from the assumptions that underly OLS affect the results of our analysis?

We ordinarily break that down in several steps.

First, in the longitudinal analysis literature, it is common to assume the variance/covariance matrix is "block diagonal". That means there can be correlations of error terms within each cluster, but the observations of the clusters are not influenced by events in other clusters. In the i'th unit, at the j'th time, the variance is σ^2_{ij} and the covariance of errors within that unit at times s and t is $\sigma^2_{ist} = E(e_{is}, e_{it})$.

$$Var(e) = \begin{bmatrix}
\sigma^2_{11} & \dots & \sigma^2_{11T} & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 \\
\sigma^2_{112} & \dots & \sigma_{12T} & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 \\
 & \dots & \vdots & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 \\
\sigma^2_{11T} & \dots & \sigma^2_{1TT} & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 \\
0 & 0 & 0 & \sigma^2_{21} & \dots & \sigma^2_{21T} & 0 & 0 & 0 & & 0 & 0 \\
0 & 0 & 0 & \vdots & & \vdots & 0 & 0 & 0 & & 0 & 0 \\
0 & 0 & 0 & \sigma^2_{21T} & \dots & \sigma^2_{2T} & 0 & 0 & 0 & & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{31} & \dots & \sigma^2_{31T} & & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \vdots & \ddots & \vdots & & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{31T} & & \sigma_{3T} & & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \vdots & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \sigma^2_{N(T-1)} & \sigma^2_{N(T-1)T} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \sigma^2_{N(T-1)T} & \sigma^2_{NT}
\end{bmatrix} \tag{1}$$

Sometimes a symbol such as Ω is used instead of $Var(e)$.

It is often easier to think of this matrix as a block-diagonal matrix in which each unit's error terms are intercorrelated according to the variance/covariance matrix V_j, which is TxT.

$$Var(e) = \Omega = \begin{bmatrix}
V_1 & & & & 0 \\
 & V_2 & & & \\
 & & V_3 & & \\
 & & & \ddots & \\
0 & & & & V_N
\end{bmatrix} \tag{2}$$

Second, the next big simplifying assumption is often that these submatrices are of a common sort. They may be assumed to be identical, or to differ according to just one or two coefficients. Let's suppose that the blocks have exactly the SAME correlation structure, V.

$$\Omega = \begin{bmatrix}
V & & & & 0 \\
 & V & & & \\
 & & V & & \\
 & & & \ddots & \\
 & & & & V \\
0 & & & & V
\end{bmatrix} \tag{3}$$

2 Detours into matrix algebra

The data matrix X is a "stack" of smaller matrices, one for each cluster. Suppose there is an intercept and 3 variables to be estimated:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_{N-1} \\ X_N \end{bmatrix} = \begin{bmatrix} 1 & x1_{11} & x2_{11} & x3_{11} \\ 1 & x1_{12} & x2_{12} & x3_{12} \\ 1 & x1_{1T} & x2_{1T} & x3_{1T} \\ 1 & x1_{21} & x2_{21} & x3_{21} \\ 1 & x1_{22} & x2_{22} & x3_{22} \\ 1 & x1_{23} & x2_{23} & x3_{23} \\ 1 & x1_{31} & x2_{31} & x3_{31} \\ 1 & & & \\ 1 & & & \\ 1 & x1_{N(T-1)} & x2_{N(T-1)} & x3_{N(T-1)} \\ 1 & x1_{NT} & x2_{NT} & x3_{NT} \end{bmatrix}$$

Similarly, the vector of observations:

$$y = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{2T} \\ \vdots \\ y_{N1} \\ \vdots \\ y_{NT} \end{bmatrix}$$

Because the data is so obviously separable into clumps, there are some special elements from matrix algebra that arise.

2.1 Kronecker product

The expression (3) can be written more compactly if we use the notation of the "Kronecker product" $\otimes$.

$$I \otimes V$$

The Kronecker product means that one takes each term in the first matrix and multiplies it by the ENTIRE second matrix and then puts the result in place of the element of the first matrix. Since I is the identity matrix

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

It is quite easy to imagine what $\otimes$ is doing. Take each element of the I matrix, multiply it by the matrix V. The result is either $1 * V = V$ or $0 * V = 0$. When that result is put into the identity matrix in place of the 0 or the 1, then the result is 3 in the short form or the larger thing in the matrix above it.

One often sees the Kronecker product defined as:

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & & & \\ a_{m1}B & a_{m2}B & & a_{mn}B \end{bmatrix}$$

Where A is an $m \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Some authors use the Kronecker product a lot, others use the nature of the block-diagonal matrix in order to simplify their findings.

2.2 The inverse of a block-diagonal matrix.

The inverse of a block-diagonal matrix is made-up of the inverses of the individual cluster matrices:

$$\Omega^{-1} = \begin{bmatrix} V_1^{-1} & 0 & 0 & 0 & 0 \\ 0 & V_2^{-1} & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & V_{N-1}^{-1} & 0 \\ 0 & 0 & 0 & 0 & V_N^{-1} \end{bmatrix}$$

2.3 The cross-product of a block-diagonal matrix

The product $X'X$ is the block-diagonal matrix made up of products of the individual cluster matrices.

$$X'X = \begin{bmatrix} X_1'X_1 & 0 & 0 & 0 & 0 \\ 0 & X_2'X_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & X_{N-1}'X_{N-1} & 0 \\ 0 & 0 & 0 & 0 & X_N'X_N \end{bmatrix}$$

2.4 Behold: The information matrix.

$$X'\Omega^{-1}X = \sum_{i=1}^{N} X_i'V_i^{-1}X_i$$

3 Recall OLS

Assume:

1. $\hat{y} = Xb + e$
2. Homoskedasticity and no autocorrelation: $Var(e) = \sigma^2 I$
3. $E(e) = 0$

which implies $\hat{y} = X\hat{b}$

The sum of squared errors:
$SS(\hat{b}) = (y - \hat{y})'(y - \hat{y}) = (y - X\hat{b})(y - X\hat{b})$
From elementary calculus, $\frac{\partial}{\partial b}(bx) = x$. The same is true of matrices,

$$\frac{\partial}{\partial \hat{b}}(X\hat{b}) = X \tag{4}$$

The derivative of SS wrt $\hat{b}$ is

$$\frac{\partial SS}{\partial \hat{b}} = -2\left[\frac{\partial \hat{y}}{\partial b}\right]'(y - \hat{y}) = -2X'(y - X\hat{b}) = 0 \tag{5}$$

This uses the fact stated in 4.

The constant -2 gets "divided away". As a result, the first order condition is the same as the "score equation" in maximum likelihood:

$$X'(y - \hat{y}) = X'(y - X\hat{b}) = 0 \tag{6}$$

The solution for the best OLS estimator:
$\hat{b} = (X'X)^{-1}X'y$
and the variance/covariance matrix of the b's is estimated by
$Var(\hat{b}) = \sigma^2(X'X)^{-1}$

If you don't know σ^2 estimate it from the residuals on the regression line. The Mean Square Error is estimate of σ^2
$\hat{\sigma}^2 = \frac{e'*e}{T-M} = \frac{(sum.of.squared.resituals)}{(N.of.cases)-(N.of.elements.in.b)}$

4 Recall GLS/WLS

Check my GLS handout. Basically:

If your assumptions about Ω are violated, there is a fix.

Recall that, unless you employ a fix, the estimates of $\hat{b}$ are inefficient and the estimates of $Var(\hat{b})$ are simply wrong.

If we know what Ω is, the solution from WLS/GLS is a weighted regression. Use Ω in the Sum of Squared formula to weight observations so that "high variance" cases have less weight.

$$SS(\hat{b}) = (y - \hat{y})\Omega^{-1}(y - \hat{y}) \tag{7}$$

The GLS equivalent of the OLS equation 5 is:

$$\frac{\partial SS}{\partial \hat{b}} = -2\left[\frac{\partial \hat{y}}{\partial \hat{b}}\right]'\Omega^{-1}(y - \hat{y}) = -2X'\Omega^{-1}(y - X\hat{b}) = 0 \tag{8}$$

This reduces to:

$$\left[\frac{\partial \hat{y}}{\partial \hat{b}}\right]' \Omega^{-1}(y - \hat{y}) = X'\Omega^{-1}(y - X\hat{b}) = 0 \tag{9}$$

The estimator ends up being a close parallel to the OLS solution for $\hat{b}$:
$\hat{b} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$
and the variance of $\hat{b}$ is:
$Var(\hat{b}) = (X'\Omega^{-1}X)^{-1}$

5 Recall "robust" variance matrix estimates

The White-Huber estimate of the standard error is a "heteroskedasticity consistent" estimate of the standard error of $\hat{b}$. Look at the end of my GLS handout, where the robust estimate is stated in equation 15.

$$Var_{hc0}(b) = (X'X)^{-1}X'\{\widehat{Var[e]}\}X(X'X)^{-1} \tag{10}$$

$$Var_{hc0}(b) = (X'X)^{-1}X'\{diag[\hat{e}\cdot\hat{e}']\}X(X'X)^{-1} \tag{11}$$

The middle part is an "empirical estimate" of the variance of the error terms.

$$\widehat{Var(e)} = diag[\hat{e}\cdot\hat{e}'] = \begin{bmatrix} \hat{e}_1^2 & & & & & 0 \\ & \hat{e}_2^2 & & & & \\ & & \hat{e}_3^2 & & & \\ & & & \ddots & & \\ & & & & \hat{e}_{N-1}^2 & \\ 0 & & & & & \hat{e}_N^2 \end{bmatrix}$$

In White's original paper, he assumes there is no autocorrelation, we are looking only at the main diagonal.

6 Normally distributed variables, PCSE, and what not

There is a separate handout called CXTS-PSCE that details the meaning and importance of the "panel corrected standard error" as proposed in Beck & Katz (American Political Science Review, 1995)

7 Longitudinal with Normal Data: FGLS returns?

Suppose we assume the error term satisfies the core assumption of longitudinal data analysis, which is that the observations within clusters are correlated with each other, but are not affected by observations in other clusters.

7.1 As the Song says, "If you knew Omega, like I know Omega, Oh, Oh, Oh what a Matrix..."

The GLS and the Maximum Likelihood approaches lead to the same conclusion. Suppose Ω is as specified in expression 2.

Remember we are thinking of y as a "stacked column" of observations on clusters, y_1, y_2, ..., y_N. And X is a stacked matrix of observations on clusters, and so forth.

The Score equations are

$$\frac{\partial lnL}{\partial \hat{b}} = X'\Omega^{-1}(y - X\hat{b})$$

Because we assume the clusters are separable, this can be written as a sum of within-cluster results:

$$= \sum_{i=1}^{N} X_i' V_i^{-1}(y_i - X_i\hat{b}) = 0$$

Solve that for $\hat{b}$, figure out an estimator for $Var(\hat{b})$, and all the work is done.

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

$$= (\sum_{i=1}^{N} X_i' V_i^{-1} X_i)^{-1}(\sum_{i=1}^{N} X_i' V_i^{-1} y_i)$$

$$Var(\hat{b}) = (X'\Omega^{-1}X)^{-1} = \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1}$$

7.2 FGLS: If you don't know Omega

I'm looking at Dobson (2002, p. 200). Repeat:

Estimate the parameters $\hat{b}$

calculate the residuals,

estimate $\hat{\Omega}$

recalculate the $\hat{b}$.

There are some standard suggestions for "working models" of V_i, such as

1. Exchangeable

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho & \rho & \rho \\ \rho & 1 & & & \rho & \rho \\ & & 1 & & & \rho \\ \vdots & & & \ddots & & \\ \rho & \rho & & & & \\ \rho & \rho & \rho & & & 1 \end{bmatrix}$$

2. AR(1)

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & & & \rho^{T-3} & \rho^{T-2} \\ & & 1 & & & \rho^{T-3} \\ \vdots & & & \ddots & & \\ \rho^{T-2} & \rho^{T-3} & & & & \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & & & 1 \end{bmatrix}$$

3. Unstructured

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1(T-2)} & \rho_{1(T-1)} & \rho_{1T} \\ \rho_{21} & 1 & & & \rho_{2(T-1)} & \rho_{2(T)} \\ & & 1 & & & \rho_{3(T)} \\ \vdots & & & \ddots & & \\ \rho_{(T-1)1} & \rho_{(T-1)2} & & & & \\ \rho_{T1} & \rho_{(T-1)1} & \rho_{(T-2)1} & & & 1 \end{bmatrix}$$

7.3 Watch out for those nonrobust standard errors!

Liang and Zeger (1986) pioneered the quasi-likelihood/GEE approach to longitudinal data analysis. Their robust estimator of the variance/covariance matrix of $\hat{b}$ is stated on their p. 15. If one puts the linear model with Normally distributed error term into that framework, then the Z&L expression simplifies radically (because $\theta_{ij} = \eta_{ij}$, so$\Delta_i = I$). The simplified version for Normally distributed dependent variables with the identity link is stated in Dobson, 2002, p. 200

$$V(\hat{b}) = \left(X'\hat{\Omega}X\right)^{-1} \left(\sum_{i=1}^{N} X_i' V_i^{-1} \left((y_i - X_i\hat{b})(y_i - X_i\hat{b})'\right) V_i^{-1} X_i\right) \left(X'\hat{V}X\right)$$

8 What if there are "random effects" at the unit level?

Suppose there is some effect at the level of the unit.

$$y_{it} = c + \gamma_i + X_{it}b + e_{it}$$

There are many different names for models that attempt to take this into account. Here are some:

- variance components model
- mixed model (because some coefficients are random and some are fixed)
- hierarchical linear model

Suppose that the random effect γ_i is Normally distributed with variance σ_γ^2.

From an econometric standpoint, this turns into a problem of heteroskedasticity, because the unobserved unit-level error term γ_i is dissolved into the unobserved individual level error term e_{it}. A GLS approach can be used to deal with the heteroskedasticity.

There is a separate handout CXTS-ECM that investigates these as error components models.

In the longitudinal data analysis literature, such as the Diggle, et al book (which is authoritative), they treat this as a maximum likelilhood problem, one which is thought of as a 'conditional' problem. It is conditional in the sense that we would first like to know the value of γ_i and then we think of the impact of $X_{it}b$ as "above and beyond" that random intercept. That's the sense in which our understanding of y_{it} is conditional on γ_i.

I'm struck by the wide variety of terminology and approaches to these models.

9 What if there are Fixed Effects at the unit level

The famous "least squares dummy variables" model. Blech!

10 What if the dependent variable is not normal, or the link function is not the identity function, or both?

Recall the GLM. It deals with nonNormal variables and wild link functions.

But it does not directly translate to deal with a longitudinal data exercise.

Part of the problem is that, up to this point, our discussion of the longitudinal model has followed the old econometric tradition of talking about the distribution of an "error term."

In most GLM applications, there is no such thing as an "error term." In the GLM, we talk directly about y being distributed as Poisson or Gamma.

Since the GLM does not have an error term, it is not obvious where one should fit in intercorrelated errors!

Liang and Zeger (1986) proposed a modeling strategy that they called GEE, Generalized Estimating Equations. GEE is an extension of the "quasi-likelihood" approach to estimation. Suppose for each cluster there is a covariance matrix for the observed values of y.

I'm writing the GEE details down in a separate handout, but I want to point out the continuity with the GLS model.

The GEE is defined as the solution to an equation that looks like a hybrid between the Score equation from the Quasi-likelihood model and the Score from the GLS in 8. Suppose the estimated mean vector is $\hat{\mu}$.

$$\left[\frac{\partial \hat{\mu}}{\partial \hat{b}}\right]' \Omega^{-1}(y - \hat{\mu}) = 0$$

That's just like the GLS score equation, because it has a weight matrix Ω in the middle. But it is different from GLS $\left[\frac{\partial \hat{\mu}}{\partial \hat{b}}\right]'$ does not resolve to a simple thing like X'.

And its different from the GLM, because it has Ω^{-1}.