Cross Sectional Time Series: The Normal Model and Panel Corrected Standard Errors

Paul Johnson <pauljohn@ku.edu>

5th April 2004

The Beck & Katz (APSR 1995) is extremely widely cited and in case you deal with panel data, political science readers will expect you to be familiar with it.

The gist of the matter is this. The econometric theory of GLS typically assumes you know Ω or can approximate it. But the approximation procedures that have been proposed are under attack in the Beck & Katz paper.

1 Correlations across units, heterogeneity across units

I hasten to point out that this is not exactly relevant to the longitudinal data problem because it is partly aimed at estimating correlations “across units”. That’s what they call the “contemporaneous correlation” problem. From Beck and Katz, (p. 645):

Panel Heteroscedasticity means that the variance of the error term within a cluster is constant, but it varies across clusters. $E(e_{it}^2) = E(e_{is}^2) = \sigma_i^2$ but it varies across units: $E(e_{it}^2) \neq E(e_{jt}^2)$,

Contemporaneously Correlated Errors. $E(e_{is}e_{js}) = E(e_{it}e_{jt}) = \sigma_{ij} \neq 0$, but there is no correlation across time $E(e_{is}e_{jt}) = 0$.

So, in contrast to the standard longitudinal data problem in which the separate units are uncorrelated with each other, we have in mind the additional problem of shared error effects across units.

To focus on the PCSE, we act as if there no autocorrelation within a unit or that it has been removed by a statistical correction.

Suppose there are 3 observations per unit. Then the error variance matrix would look like this.

$$Var(e) = \Omega = \begin{bmatrix}
\sigma_1^2 & 0 & 0 & \sigma_{12} & 0 & 0 & & \sigma_{1N} & 0 & 0 \\
0 & \sigma_1^2 & 0 & 0 & \sigma_{12} & 0 & & 0 & \sigma_{1N} & 0 \\
0 & 0 & \sigma_1^2 & 0 & 0 & \sigma_{12} & & 0 & 0 & \sigma_{1N} \\
\sigma_{12} & 0 & 0 & \sigma_2^2 & 0 & 0 & & \sigma_{2N} & 0 & 0 \\
0 & \sigma_{12} & 0 & 0 & \sigma_2^2 & 0 & & 0 & \sigma_{2N} & 0 \\
0 & 0 & \sigma_{12} & 0 & 0 & \sigma_2^2 & \cdots & 0 & 0 & \sigma_{2N} \\
 & & \vdots & & & \vdots & \ddots & & & \\
\sigma_{1N} & 0 & 0 & \sigma_{2N} & 0 & 0 & & \sigma_N^2 & 0 & 0 \\
0 & \sigma_{1N} & 0 & 0 & \sigma_{2N} & 0 & & 0 & \sigma_N^2 & 0 \\
0 & 0 & \sigma_{1N} & 0 & 0 & \sigma_{2N} & & 0 & 0 & \sigma_N^2
\end{bmatrix}$$

When Ω is written in this way, it is assumed that, if there is autocorrelation, it has already been eliminated from the data. There may also be autoregressive error within each separate unit. Following the econometric literature that preceded their effort, Beck and Katz suppose that the autocorrelated errors follow an AR(1) pattern:

$$e_{it} = \rho e_{i,t-1} + v_{it}$$

where v_{it} are nice, well behaved Normal error terms with mean 0 and fixed variance.

This AR(1) pattern is not the primary focus of the analysis. It is“purged” in a first stage of analysis in the usual econometric way, resulting in observations that are correlated as described in Ω above.

2 Big, Complicated FGLS mess

Parks’s FGLS (Feasible Generalized Least Squares) is one way that was proposed to deal with CXTS data. One estimates with OLS, then uses the residuals to calculate autocorrelation and heteroskedasticity, then estimate again with GLS. The procedure may be repeated many times, until the estimates of the b’s converge to a fixed number.

When there is autoregression to be considered, it adds another layer. Park’s method is one way of attacking this problem with CSTS data.

1. estimate by OLS

2. Use residuals to estimate AR(1) models.

3. Use AR(1) models to adjust data and estimate again. (see standard stats books under corrections for AR(1), such as Prais-Winsten estimators).

4. Use residuals to estimate cross-correlation across units (spatial autocorrelation)

5. Use results from 4 to fill in more values of Ω and estimate model with GLS again.

3 FGLS dangers

The multi-stage Parks method has dangers. Beck and Katz have 2 major arguments.

1. The estimates of the b’s are more inefficient (higher variance) than OLS, and

2. The estimates of the variances of the b’s from the last stage GLS are biased downwards. This happens because the estimate of Ω is never exactly equal to Ω.

The basic argument is that the process of repeatedly estimating $b$and Ω can “compound” inaccuracy in the standard error of $\hat{b}$ estimates. The estimates of the standard errors don’t take into account uncertainty of the Ω estimates, but rather just take the estimates and plug them in!

4 Beck and Katz recommend

1. Use OLS to estimate the b’s, but
2. correct the estimates of the standard errors of the b’s–use “panel corrected standard errors.”
3. If you suspect there is AR in the errors, they say it should be corrected by one of the simple AR(1) adjustments, such as the Prais-Winsten estimator. After that correction is applied, the errors should follow the pattern assumed in Ω and the PCSE can be calculated from the residuals.

The “**panel corrected standard errors**” are a “robust” estimate in the sense of White-Huber.

Monte Carlo studies by B&K support the claim that OLS estimators have lower variance than the Parks estimates, and that the Parks estimates lead to false t statistics because the Variance of b is underestimated.

If we assume the AR coefficient ρ_i is the same for all countries, then we are only estimating one coefficient. We should not give in to the temptation to estimate ρ_i separately for each unit because doing to cause us to fit N different coefficients, and the accumulated uncertainty about those estimates can be large.

Beck and Kats contend that in practice the ρ should be the same for all units. They also have Monte Carlo estimates to claim that, even if ρ is not the same across all countries, it doesn't do much harm to assume that it is in order to make these calculations.

5 The Panel Corrected Standard Error

Why not just use the Huber-White formula to calculate a robust standard error?

That throws away information! It ignores the fact that we assume there is a common variance structure within a cluster and that the intercorrelation across units follows a very specific pattern–equal covariance between any 2 units for any particular time.

So Beck and Katz propose an estimator that it pools information across clusters to estimate the error variances. The Beck & Katz panel corrected standard error is calculated in the following way.

Organize the residuals from the fitted model according to cluster, so that the residuals from the clusters are $\hat{e}_1$, $\hat{e}_2$, ... , $\hat{e}_N$. These are vectors with T elements each, and they can be grouped together as a $T \times N$ matrix (the $\hat{e}_i$ are columns):

$$E = \left[\begin{array}{ccccc} \hat{e}_1 & \hat{e}_2 & \cdots & \hat{e}_{N-1} & \hat{e}_N \end{array} \right]$$

The panel corrected variance/covariance matrix of $\hat{b}$ is

$$PCSEVar(\hat{b}) = (X'X)^{-1}X'\left(\hat{\Omega}\right)X(X'X)^{-1}$$

Note that is a "sandwich estimator" in the style of the Huber-White robust estimator. The only difference is that $\hat{\Omega}$ is estimated differently.

The only thing about all of this that is in the slightest bit complicated in calculating $\hat{\Omega}$ is the use of the "Kronecker product". Beck and Katz use the standard symbol for this $\otimes$. On p. 646 they write:.

$$\hat{\Omega} = \frac{(E'E)}{T} \otimes I \qquad (1)$$

The Kronecker product is defined in many matrix algebra books as well as in my CXTS#1 handout. Multiply a chosen term in the matrix on the left with the matrix on the right, and place the result into the matrix on the left in place of the chosen element.

To get an idea what these terms mean, consider a couple of details. The estimate of the variance of the error term for cluster 1 is

$$\hat{\sigma}_1^2 = \frac{1}{T}\{\hat{e}_{11}^2 + \hat{e}_{12}^2 + \cdots + \hat{e}_{1T}^2\}$$

In other words, it is the mean squared error for unit 1. For each individual unit, the variance of its error term is estimated as the MSE of its residuals.

In the PCSE, this estimate is assumed to apply for all time points for cluster 1. In contrast, the White-Huber robust estimator would not take account of that information. In the White-Huber robust estimator, each observation would be treated as a separate thing. Where the PCSE would lead to the estimates for cluster 1 that look like this:

$$\hat{V}_1^{PCSE} = \begin{bmatrix} \hat{\sigma}_1^2 & 0 & 0 \\ 0 & \hat{\sigma}_1^2 & 0 \\ 0 & 0 & \hat{\sigma}_1^2 \end{bmatrix} \tag{2}$$

the White-Huber estimator would have the squared residuals along the diagonal.

$$\hat{V}_1^{hc0} = \begin{bmatrix} \hat{e}_{11}^2 & 0 & 0 \\ 0 & \hat{e}_{12}^2 & 0 \\ 0 & 0 & \hat{e}_{13}^2 \end{bmatrix}$$

The estimates in the diagonal of the PCSE are more precise because they average together several observations-worth of error terms to estimate the error variance. The White-Huber approach does not take advantage of this pooling because it does not take clustering into account.

The covariance across units in $\hat{\Omega}$ is estimated by taking the residuals from the two units and calculating their cross product. For example, consider a cross correlation, say between units 5 and 9:

$$\sigma_{59} = \frac{1}{T}\{\hat{e}_{51}\hat{e}_{91} + \hat{e}_{52}\hat{e}_{92} + \cdots + \hat{e}_{5T}\hat{e}_{9T}\}$$

This is combining the residuals across all time points because a central assumption in the model is that the "contemporaneous correlation across units" follows a fixed pattern. This approach works on the premise that the intercorrelation between units is the same for all time points, so it just averages the covariances across all time points.

6 You can calculate the PCSE for yourself. Easily.

I recently had occasion to calculate the PCSE for a regression model in R and was startled by how simple it is to calculate. If you examine footnote 15 of the original Beck and Katz paper, you find it is really quite simple. Here's some R code. The pound sign indicates comments:

```
# testmodel is a fitted ols equation
#get the long column of residuals:
resids <- residuals(testmodel)
# E is the T x N matrix of residuals, T observations for each STATE (my grouping variable)
E <- as.matrix(unstack(resids, form=resids~STATE))
# $\Sigma = (1/T)(E'E)$ an NxN is empirical covariance matrix,
SIGMA <- (1/nrow(E))*(t(E) %*% E)
# $\hat{\Omega}$ is the matrix (NT x NT) of estimated error correlations
OMEGA <- SIGMA %x% diag(x=1, nrow=nrow(E), ncol=nrow(E))
# %x% is the R notation for the Kronecker product $\otimes$.
# Next, grab "model" data
X <- model.matrix(testmodel)
# We always seem to need $(X'X)^{-1}$
XPRIMEXINV <- solve(t(X) %*% X)
# PSCECOV(b) = $(X'X)^{-1}X'\,\hat{\Omega}\,X(X'X)^{-1}$
PCSECOVB <-XPRIMEXINV %*% (t(X) %*% OMEGA %*% X ) %*% XPRIMEXINV
```

```
# The standard errors are $\sqrt{diag(PCSECOV(\hat{b})}$
PCSEB <- sqrt(diag(PCSECOVB))
# Here's a column of t statistics
coef(testmodel)/PCSEB
```

An R printout showing these calculations and comparing them against the uncorrected standard errors is found in the file PSCEExample.txt

7 Different from robust estimates as found in GEE?

It is widely agreed that after a model is fitted, one should be conscious of violations in the assumptions when calculating standard errors and doing significance tests. If you use an approach like GEE, most programs will report both an "model based" standard error, one that uses the assumption that the covariance of the observations is specified correctly, and a "robust" standard error.

The robust standard errors used in GEE are not exactly the same as the ones used in PCSE because GEE (and longitudinal models generally) assume the units do not influence each other. But the robust estimator for GEE is extremely similar to the PCSE because both are "information sandwich" estimators.

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 Normally distributed error term of B&K into that framework, then the Z&L expression simplifies radically (because $\theta_{ij} = \eta_{ij}$, so$\Delta_i = I$). The simplified version of the robust standard error for Normally distributed dependent variables is stated in Dobson, 2002, p. 200

$$robustVar(\hat{b}) = \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\Omega^{-1}\left(\widehat{Var(e)}\right)\Omega^{-1}X\left(X'\hat{\Omega}^{-1}X\right)^{-1}$$

The information matrix is $(X'\Omega^{-1}X)$. Notice how this really is an information sandwich.

$(X'X)$ is a $p \times p$ matrix, where p is the number of columns of X.

$(X'\Omega^{-1}X)$ is also a $p \times p$ matrix.

The observed residuals from the regression are used to calculate the middle element. It can be viewed as a sum across clusters

$$robustVar(\hat{b}) = \left(X'\hat{\Omega}^{-1}X\right)^{-1}\left(\sum_{i=1}^{N} X_i'\hat{V}_i^{-1}\,\hat{e}_i\hat{e}_i'\,\hat{V}_i^{-1}X_i\right)\left(X'\hat{\Omega}^{-1}X\right)^{-1}$$

The middle part appears as a sum because we are "assuming away" the contemporaneous correlations across units.

Focus on the middle part, say, for cluster 1:

$$X_1'\hat{V}_1^{-1}\,\hat{e}_1\hat{e}_1'\,\hat{V}_1^{-1}X_1$$

$$= X_1'\hat{V}_1^{-1}\begin{bmatrix}\hat{e}_{11}\\ \hat{e}_{12}\\ \hat{e}_{13}\end{bmatrix}\begin{bmatrix}\hat{e}_{11} & \hat{e}_{12} & \hat{e}_{13}\end{bmatrix}\hat{V}_1^{-1}X_1$$

$$= X_1'\hat{V}_1^{-1}\begin{bmatrix}\hat{e}_{11}^2 & \hat{e}_{11}\hat{e}_{12} & \hat{e}_{11}\hat{e}_{13}\\ \hat{e}_{11}\hat{e}_{12} & \hat{e}_{12}^2 & \hat{e}_{12}\hat{e}_{13}\\ \hat{e}_{11}\hat{e}_{13} & \hat{e}_{12}\hat{e}_{13} & \hat{e}_{13}^2\end{bmatrix}\hat{V}_1^{-1}X_1$$

In a sense, the GEE robust SE is both more and less robust than the PCSE, at the same time. It is more robust in two senses. First, it does not assume that the variance of the error term is the same for all time points. Second, it allows the error term correlations within a cluster to take on any observed value. Unlike the PCSE, it does not set all of those diagonal elements to a constant $\hat{\sigma}_1$ and restrict the off diagonal elements to 0, as does the PCSE in expression 2.

On the other hand, the GEE SE sets all covariances across units to 0, but the PCSE estimates them. So there's a sense in which the GEE is less general.