CXTS (Panel) Notes on Error Correction Models.

Paul Johnson <pauljohn@ku.edu>

5th April 2004

1 The Panel Data Problem

We gather a few (say, 5) observations on each of 50 states. Should we do 5 sets of research, one for each "snapshot"?

2 Don't just "pool" all the observations.

You might stack all the observations into one big dataset and then proceed with regression as if there were 200 separate observations.

That's bad because it:

1. Ignores substantively significant elements you might want to isolate (effect of time or location)
2. Ignores possibility of heteroskedasticity
3. Ignores possibility of autoregression

3 Dummy Variables to the rescue

We are VERY FAMILIAR with the idea of using dummy variables as intercept and slope shifters.

The least squares dummy variable (LSDV) approach is to customize the regression model by adding dummy variables for the time and location of observations, to find out if observations for New York are different from California, or if observations in 1980 are different from observations in 1982.

The big problem: too many dummy variables floating around! If you have 50 states, you need 49 dummy variables. If you want to cross-check time effects, then add more.

And still that does not solve the fundamental issues of heteroskedasticity, autocorrelation.

4 Background: Random Effects Models.

Sometime after I took my last regression class (1983) and now, a new rage has swept through the regression community: random effects models. These were discussed before 1983, but only in very limited contexts and esoteric books. Now these ideas have become more well known and have found their way into stats books, such as Gujarati Chapter 16.

4.1 Randomly varying slope:

Here's the idea. Suppose that β is not a constant. Rather, it is a random variable. So we would write β_i. How odd looking that is. And a regression model with a random coefficient would be

$$y_i = \beta_0 + \beta_{1i} x_i + u_i \tag{1}$$

Then try this standard "sneaky trick" that is loved by econometrics professors far and wide. Suppose that the coefficient β_{1i} really has 2 parts. It has a "fixed part" β_1 as well as an additive "error term" of its own:

$$\beta_{1i} = \beta_1 + u_i \tag{2}$$

It is very important to add some conditions. We need the random element u_i to be well behaved. It needs to have an expected value of 0 and a constant variance. If so, then

$$E(\beta_{1i}) = \beta_1$$

If that is correct–so the slope is just a constant plus or minus an individual level error, then you can insert 2 into 1 and you end up with:

$$y_i = \beta_0 + \beta_1 x_i + u_i + e_i * x_i$$

This appears to be a relatively simple problem of heteroskedasticity.

Interpretation: the regression tools we have used so far can be extended to include "random effects." The estimates of the b's from ordinary least squares can be thought of as estimates of the average effects of randomly varying coefficients.

4.2 Randomly varying intercept

Its even simpler if the intercept is the random coefficient. If

$$\beta_{0i} = \beta_0 + e_i \tag{3}$$

then in the regression model

$$y_i = \beta_{0i} + \beta_1 x_i + u_i \tag{4}$$

we would simply insert 3 into 4 and get:

$$y_i = \beta_0 + \beta_1 x_i + e_i + u_i$$

If the two unobserved errors are just sitting there by themselves, it must mean we should use ordinary least square.

The estimate of $\beta_0$is an estimate of the average of the intercepts across units.

4.3 Terminology: Mixed Models

A mixed model is one in which some coefficients are random and others are fixed.

5 Error Components (Gujarati, p. 649) approach to Panel Data

5.1 Randomly varying intercept across units.

A panel model in which we hypothesize there is a randomly varying intercept across units is called an *error components model*.

In a panel study context, the idea is that each unit is randomly assigned an intercept, and then at each time point there is a second random error.

N: number of units observed

T: number of observations per unit

N*T = total number of observations

i indexes units

t indexes times

$$y_{it} = \beta_{0i} + \beta_1 x_i + u_{it}$$

and if you insert 2 into this, you have:

$$y_{it} = \beta_0 + \beta_1 x_i + e_{it} + u_i$$

We have already imposed a lot of regularity on this problem. We wave the mathemagical pen to assert that u_{it} is uncorrelated across time and space. And further e_i is uncorrelated with u_{it}. So the error term has the expected value of 0:

$$E(e_{it} + u_i) = E(e_{it}) + E(u_i) = 0$$

and the variance is:

$$V(e_{it} + u_i) = V(e_{it}) + V(u_i) + 2Cov(e_{it}, u_i)$$

but we've already used the heavy hand to assert that the e_{it}'s are uncorrelated with the u_i's, so this reduces to:

$$V(e_{it} + u_i) = V(e_{it}) + V(u_i)$$

That's homoskedastic!

But Gujarati notes on p. 648 there is autocorrelation of a particular sort. Take a look at the combined error term, which he calls w_{it}

$$w_{it} = e_{it} + u_i$$

Note that, at each time point, there are two parts of random variation, but really there is only one. The random number "u_i" is the same for all observations on a given unit. That means there is a common factor influencing the error term across time within a single unit.

Gujarati claims (p. 648) that the correlation coefficient between two observations at times s and t on a unit i, w_{it} and w_{is}, have the following correlation:

$$r_{w_{it}w_{is}} = \frac{V(e_{it})}{V(e_{it}) + V(u_i)} \tag{5}$$

5.2 Digression on where formula 5 comes from

I had to stare at that a long time to remember where it comes from. It is written down in Greene's Econometric Analysis 5th ed, p. 294, where he observes that:

$$E(w_{it}^2|x) \quad = \quad V(e_{it}) + V(u_i) = E(e_{it}^2) + E(u_i^2)$$

Greene adds the conditional notation |x because he wants to remind you that we are looking at a particular given input variable x (not a random variable x).

and within a unit, the covariance between observations is:

$$E(w_{it} * w_{is}|x) = V(e_{it})$$

This is so because the intercept is not varying within the unit, so $V(u_i|x) = 0$ and there is no covariance between e and u.

$$E(w_{it} * w_{is}|x) = E((e_{it} + u_i)(e_{it} + u_i)|x)$$

$$= E(e_{it}^2|x) + E(u_i^2|x) + 2E(e_{it}u_i|x)$$

$$= E(e_{it}^2|x)$$

You can get from there to the correlation coefficient WITHIN a unit by remembering the formula for the correlation between 2 variables x and y is:

$$r_{x,y} = \frac{Cov(x,y)}{Var(x)Var(y)}$$

and

$$Cov(x,y) = E(x - E(x)) * E(y - E(y))$$

5.3 So the error term's covariance matrix is...

Within a given unit, the errors are intercorrelated as described in the previous section. The covariance matrix looks like this if there are, say, 6 observations for the unit:

$$\Sigma = \begin{bmatrix} V(e)+V(u) & V(e) & V(e) & V(e) & V(e) & V(e) \\ V(e) & V(e)+V(u) & V(e) & V(e) & V(e) & V(e) \\ V(e) & V(e) & V(e)+V(u) & V(e) & V(e) & V(e) \\ V(e) & V(e) & V(e) & V(e)+V(u) & V(e) & V(e) \\ V(e) & V(e) & V(e) & V(e) & V(e) & V(e)+V(u) \end{bmatrix}$$

And then the "BIG" covariance matrix for the whole model would have these Σ things surrounded by 0's:

$$\Omega = \begin{bmatrix} \Sigma & & & & & & & & 0 \\ 0 & \Sigma & & & & & & & \\ & & \Sigma & & & & \ldots & & \\ & & & \Sigma & 0 & & & & \\ & & & & \Sigma & & & & \\ & & & & & \Sigma & & & 0 \\ & & & & & & \Sigma & & 0 \\ 0 & & & & & & & \Sigma & 0 \\ 0 & 0 & & & & & & & \Sigma \end{bmatrix}$$

6 Separate route of analysis through Maximum Likelihood (or REML)

In the Gujarati or other econometrics-style treatments, we are pursuing GLS and fiddling with the var/covar matrix of the error term.

If you go read some of the writings of statisticians or biostatisticians on mixed models, you find the emphasis is rather different. Consider Diggle, et al, or Pinheiro and Bates.

The Maximum Likelihood approach is arrived at by assuming that the random effect follows a Normal (Gaussian) distribution.

The REML (Restriced Maximum Likelihood) approach, which is the default in the lme package of Pinheiro and Bates. REML is strongly advocated by Diggle, et al.

Will write more later...