Appendix B

Corrected Covariance Matrices

In some cases, it won't be possible to do a GLS correction for serial correlation or cross section dependence, either because of a problem with pre-determined but not exogenous regressors (for instance, in multi-step predictability regressions) or simply because of a lack of data in one dimension. The following assumes that we fall back on using OLS, though the same ideas apply to other forms of regressions.

If we run an OLS regression on a panel data set, we get the following as the standard asymptotic breakdown:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{it} X'_{it} X_{it} \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{it} X_{it} u_{it} \right)$$

where n is the total number of data points. As usual, the first factor will be assumed to converge in probability to a fixed matrix, and the sample average serves as an estimator for it. It's the second factor which doesn't fit into any of the forms given in Hansen (1982) since, with panel data, it involves N separate time series, rather than just 1. However, the same ideas used there can be applied.

In general, we need to compute:

$$\text{cov} \left\{ \sum_{i,t} \mathbf{X}_{it} u_{it} \right\} = E \sum_{i,j,t,s} \mathbf{X}'_{it} u_{it} u_{js} \mathbf{X}_{js}$$

The sample analogue of this isn't a feasible estimator since it's a rank one matrix as the outer product of the single sum

$$\sum_{i,t} \mathbf{X}_{it} u_{it}$$

We need to make some assumptions which will reduce the freedom of the relationship.

Time Clustered Standard Errors

If we assume (conditional on X) that:

$$E u_{it} u_{js} = 0 \text{ if } t \neq s$$

that is, no time dependence (which rules out individual effects, since they create correlation between u_{it} and u_{is} for all t and s), then

$$E\sum_t \sum_{i,j} \mathbf{X}_{it}' u_{it} u_{jt} \mathbf{X}_{jt} = E\sum_t \left(\sum_i \mathbf{X}_{it} u_{it}\right)' \left(\sum_i \mathbf{X}_{it} u_{it}\right)$$

which is the covariance matrix *clustered by time*. With RATS, you get these using the option CLUSTER=%PERIOD(T) for PANEL-dated data, or CLUSTER=*time id series* for other data. Note that this will be singular if the number of time periods is less than the number of regressors since it's the sum of T rank one matrices.

Panel Corrected Standard Errors

With the similar, but stronger assumption that:

$$Eu_{it}u_{js} = \begin{cases} 0 & \text{if } t \neq s \\ \sigma_{ij} & \text{if } t = s \end{cases}$$

we get

$$E\sum_t \sum_{i,j} \mathbf{X}_{it}' u_{it} u_{jt} \mathbf{X}_{jt} = E\sum_t \sum_{i,j} \sigma_{it} \mathbf{X}_{it}' \mathbf{X}_{jt}$$

These give rise to *Panel Corrected Standard Errors* (Beck and Katz (1995)). This is an alternative to using a feasible GLS estimator. If Σ were known, GLS would be more efficient than OLS. However, if Σ isn't known, and if T isn't much larger than N, then feasible GLS can be less accurate than OLS because it weights observations using the inverse of a poorly estimated Σ matrix. While the PCSE use that same matrix, they don't invert it, and so aren't subject to the same degree of sampling error. After you estimate the equation by **LINREG**, you can use the procedure **@REGPCSE** to redo the regression with the corrected covariance matrix.

Individual Clustered Standard Errors

If instead, we assume:

$$Eu_{it}u_{js} = 0 \text{ if } i \neq j$$

that is, no correlation between individuals, we get

$$E\sum_i \sum_{t,s} \mathbf{X}_{it}' u_{it} u_{is} \mathbf{X}_{is} = E\sum_i \left(\sum_t \mathbf{X}_{it} u_{it}\right)' \left(\sum_t \mathbf{X}_{it} u_{it}\right)$$

which are (individual) clustered standard errors. These allow for arbitrary patterns of serial correlation within an individual. These can be done directly on the **LINREG** using LWINDOW=PANEL or CLUSTER=%INDIV(T) for PANEL-dated data, or CLUSTER=*individual id series* for other data. This will be singular if the number of individuals is less than the number of regressors since it's the sum of N rank one matrices.

Individual HAC Standard Errors

A somewhat sharper assumption is:

$$Eu_{it}u_{js} = 0 \text{ if } i \neq j \text{ or } |t-s| > L$$

which limits the serial correlation within an individual record to no more than L lags or leads (which again rules out individual effects). There is no correlation between individuals. A feasible estimator for this requires some type of window to prevent the covariance matrix from going non-positive definite.

$$\sum_i \sum_t \sum_{l=-L}^{L} w_l u_{it} u_{i,t-l} \mathbf{X}_{it}' \mathbf{X}_{i,t-l}$$

where the w_l are lag weights. A particular form of this would be Newey-West, which is applied individual by individual. This can be done directly on the **LINREG** using the `LAGS` and `LWINDOW` options (like `LWINDOW=NEWEY`).

Full Panel HAC Standard Errors

The most general form that can be handled is:

$$Eu_{it}u_{js} = 0 \text{ if } |t-s| > L$$

This allows for serial correlation both within and between individuals. However, again, this rules out individual effects.

$$\sum_{i,j} \sum_t \sum_{l=-L}^{L} w_l u_{it} u_{j,t-l} \mathbf{X}_{it}' \mathbf{X}_{j,t-l}$$

which is a full panel HAC correction. After you estimate the equation by **LINREG**, you can use the procedure **@REGPCSE** with the option `METHOD=PHAC` and `LAGS=`*`value of L`* to redo the regression with the corrected covariance matrix.