Panel Data[1]

We now begin the course's treatment of *panel data*. In general a panel consists of N cross sections and T time periods. If each cross section has the same number of periods and each time period has the same number of cross sections, then the panel is said to be "balanced." Otherwise, it is unbalanced. The maximum number of observations is thus NT which applies to a balanced panel, but unbalanced panels will have fewer. Panels are commonly employed in many types of applied work, including microeconomics and macroeconomics. We will consider three estimators: pooled OLS, fixed effects, and random effects. We will then discuss how to test among these options and discuss some related issues.

General Setup

Consider some variable, y_{it} that varies both cross sectionally and over time. Consider the following general regression model:

$$y_{it} = x_{it}^{'}\beta + z_{i}^{'}\alpha + g_{t}^{'}\alpha + u_{it} = x_{it}^{'}\beta + c_i + d_t + u_{it} \tag{1}$$

The matrix x consists of the regressors, excluding a constant. It is standard. α is a vector of constants. The inclusion of $z_i^{'}$ allows for cross sectional heterogeneity. Suppose, for example, that i represents different countries. France may then, for example, include unobserved (not included in x) heterogeneity that systematically affects y_{it} differently than other countries in the sample. The inclusion of $g_t^{'}$ allows for heterogeneity over time. Suppose for example that the financial crisis of 2008 is not explicitly included in x but causes y_{it} to vary systematically across countries. In this case, such heterogeneity might be significant.

Pooled OLS

Suppose that we are willing to assume that that both $z_i^{'}$ and $g_t^{'}$ are constants (normalized to one). In this case, (1) reduces to:

$$y_{it} = x_{it}^{'}\beta + \alpha + u_{it} = x_{it}^{'}\beta + c_i + d_t + u_{it} \tag{2}$$

Note that this is our standard OLS model. Pooled OLS thus ignores the potential for unobserved heterogeneity and thus ignores the panel nature of the data altogether. If the remaining

[1] These are undergraduate lecture notes. They do not represent academic work. Expect typos, sloppy formatting, and occasional (possibly stupefying) errors.

Gauss-Markov conditions hold (requiring exogeneity, stationarity, homoscedasticity,etc.), then OLS will be BLUE. Remaining violations of these conditions, however, must be corrected in the standard manner.

The key for pooled OLS to be appropriate is convincing the reader that there is no unobserved heterogeneity. Consider the following examples:

1. Our data includes the year 2008, where a rare financial crisis occurred, and y_{it} is some macroeconomic variable. We thus likely suspect that the econometric relationship is different in 2008 than in other years. If our matrix of independent variables, x includes measures of the financial crisis, then the effects of the financial crisis may be observed and we may not need to worry about their unobserved effects.

2. Suppose that our dependent variable is insurance claims on beachfront property by state. In Oregon, almost all shoreline is publicly owned. This policy would likely mean that the econometric relationship is different for the State of Oregon. If our matrix of independent variables includes a measure of publicly owned shorefront, then this heterogeneity may be observed and there is no problem. If, however, it is not, then we have a problem and pooled OLS is inappropriate.

In most cases, it is not possible to convince the reader that pooled OLS is appropriate. It is therefore mostly used as a robustness check instead of as a primary estimator.

Fixed Effects

We now make two key assumptions:

$$E[c_i|X_i] = h(X_i) \tag{3}$$

$$E[d_t|X_t] = k(X_t) \tag{4}$$

In words, the cross sectional and time effects are correlated with the right hand side variables. We can then substitute (3)-(4) into (2):

$$y_{it} = x_{it}^{'}\beta + h(X_i) + k(X_t) + [c_i - h(X_i)] + [d_t - k(X_t)] + u_{it} \tag{5}$$

The key here is that the bracket terms are, using (3) and (4), uncorrelated with x. We may then absorb them into the error term without introducing endogeneity. Doing so yields our specification:

$$y_{it} = x_{it}^{'}\beta + \alpha_i + \gamma_t + u_{it} \tag{6}$$

The fixed effect model thus includes a different estimator for each time period and each cross section. Assuming that the remaining Gauss-Markov conditions hold, then it is straightforward to estimate (6) using least squares. To do this is easy:

1. Create dummy variables for all of the time periods except one.

2. Create dummy variables for all of the cross sections except one.

3. Run OLS without a constant. Including an additional constant (either here or in #1) will create perfect multicolinearity, known as a dummy variable trap.

Because the model is just standard OLS with dummies, it is also known as the Least Squares Dummy Variable Model (LSDV). We may then be interested in comparing it to the pooled OLS model. This may be done using a simple F-test:

$$F(N-1, NT-N-T-K) = \frac{(R^2_{LSDV} - R^2_{pooled})/(N+T-1)}{(1-R^2_{LSDV})/(NT-N-T-K)} \tag{7}$$

where NT assumes a balanced panel and more generally is the number of observations and where K is the number of independent variables in x. If the resulting F value is greater than the critical value, than we reject the null that the fixed effects are jointly zero and that fixed effects are appropriate,

Another complication is that we may not always want to include fixed effects for both cross sections and time. This is often true when either N or T is small. It would be unusual, for example, to include time fixed effects in your specification when your dataset only includes 5 periods. We can formally test for whether or not to include one type of fixed effect using F tests. We could use an F test if we we wanted to test time fixed effects versus no time fixed effects but that we are sure that we want to include cross sectional fixed effects.

We can then test cross sectional fixed effects versus the pooled OLS model:

$$F(N-1, NT-N-K) = \frac{(R^2_{LSDV} - R^2_{pooled})/(N-1)}{(1-R^2_{LSDV}/(NT-N-K)} \tag{8}$$

Random Effects

The critical difference between random and fixed effects is that the former assumes that time and cross sectional effects are uncorrelated with the independent variables. In this case, we may rewrite (2) as:

$$y_{it} = x_{it}^{'}\beta + \alpha + v_t + e_i + u_{it} \tag{9}$$

The set of regressors is thus the same as pooled OLS. The complication, however, is that the three components of the error term may be correlated with each other. This is similar to the complication associated with the seemingly unrelated regressions model. We will not derive the random effects estimator. Intuitively, it optimally exploits this correlation to obtain efficient estimates.

Suppose that we run pooled OLS when we should not. There are two cases:

1. If fixed effects is the correct specification, then the exclusion of the dummy variables constitutes omitted variable bias. Our results will thus be biased.

2. If random effects are the correct specification, then pooled OLS will generally be consistent and unbiased. because information in the error terms is not exploited, however, it will be inefficient. This is similar to just running OLS when a SUR model is the correct specification.

An Example:

Suppose that we are interesting in estimating the determinates of rental housing prices in different U.S. cities. We include a set of regressors that includes income, housing prices, and a measure of rent control. Our data is annual going back to 1970. Consider the case for each of the three estimators:

1. If we believe that there is no unobserved heterogeneity than pooled OLS may be appropriate. Suppose, for example, that rents increase nationwide in a given year. If this is purely a combination of random noise and changes in the independent variables then there may be no need for time fixed effects. Likewise, if rents in a particular city are persistently high, then there is no need for cross sectional fixed effects if they are a combination of noise and the independent variables.

2. Suppose that we observe that rents in New York and San Francisco are persistently high in a way that is not random and is not explained by the independent variables. It is unlikely to be random, for example, if they are regularly several standard deviations above those predicted by

the model. If we further believe that this variation is correlated with the independent variables then we may want to include cross sectional fixed effects. This may be the case if these cities have high levels of rent control but the estimated (from pooled OLS) coefficients on rent control are not large enough to explain the variation.

Further suppose that in 2005 we observe a nationwide increase in rents. If it is too large to be random, then there may be a problem with pooled OLS. If we believe that this variation is correlated with an independent variable, then we may want time fixed effects. This may be the case if a national housing bubble (captured by housing prices) is responsible for the increase.

3. Now suppose that whatever is causing San Francisco and New York rents to be high is unrelated to any of the regressors. This could be the case if high rents are a function of quality of life that is unconnected to rent control, housing prices, or income. In this case, random effects may be appropriate if similar reasoning applies to variation over time.

Fortunately we need not rely on theory alone to choose between fixed and random effects. The *Hausman test* is the most popular test to test between fixed and random effects. Its details are highly technical. Intuitively, it does the following:

1. Under the null hypothesis, both random and fixed effects are consistent. Random effects is, however, efficient. This is the case if the dummy variables from fixed effects are close enough to zero. If we fail to reject the null, we thus run random effects.

2. We then run both specifications.

3. A test statistic (that requires sophisticated linear algebra) is then computed. The value of this test statistic is increasing in the absolute value of each dummy variable in the fixed effect specification. It is also decreasing in their standard errors.

4. If the test statistic is large enough we reject the null. Because random effects may be inconsistent, we thus run fixed effects.

The Hausman test is a good illustration of how econometricians must attempt to balance the tradeoff between efficiency and consistency. Suppose that we want to obtain the best estimate as defined by the expected squared difference between the actual value of a coefficient (which is of course unobserved) and the estimated value. This error comes from two sources, bias and variance:

$$E[(\beta - \hat{\beta})^2] = Bias^2 + Variance^2 \tag{10}$$

If the unobserved heterogeneity is uncorrelated with the independent variables, then random effects is unbiased. It is also more efficient than fixed effects because it does not include any irrelevant variables. As the correlations among the independent variables and the unobserved heterogeneity become stronger, however, random effects becomes increasingly biased. At some point, fixed effects becomes preferable.

It is important to note that sometimes a biased, but efficient, estimator may be better than an unbiased, but inefficient, estimator. In practice, any correlation is unlikely to be exactly zero. But random effects may still be preferable, even if it is biased, due to its efficiency. This is not problematic. Keep in mind that every econometric specification suffers from some omitted variable bias and is thus biased.

Other Issues

#1 Non-Stationarity: For the same reasons that non-stationarity is problematic in standard time series, it is also problematic with panel data. For short (small T) panels, this issue is often ignored in the hopes that the non-stationarity will not manifest itself over a short timeframe. For longer panels, however, it must be dealt with. The problem, however, is that a variable may be I(1) for one i, it may be I(2) for another. It is thus not obvious how to deal with this. There are two general approaches:

1. Run individual unit root tests (*e.g.* Dicky Fuller) and then difference each of the series based on the *highest* order of integration. The benefits of this strategy are that it prevents biased results. It may, however, require that we eliminate too many observations. It may also introduce additional autocorrelation from overdifferencing.

2. Use more sophisticated unit root tests that balance the concerns from #1. These, however, are beyond the scope of this class.

#2 Group Effects: Suppose that we have a panel with high N but low T. Including cross sectional fixed effects may thus be problematic because it leaves us with too few degrees of freedom. We may thus include group level fixed effects. We might for example, include a dummy for Europe, one for North America, etc. The drawback of this approach is that it is subjective how exactly we define the groups. We may also employ this method if we have low N. We could thus include a dummy for each decade, for example, instead of each year.

Suppose for example that we have two years of annual data for 100 different countries. Thus $N = 100$ and $T = 2$. We would surely not run true fixed effects, we do not have enough observations. But we might include regional dummies that serve a similar purpose. We may also include a dummy for one of the years, essentially time fixed effects.

#3 Dynamic Panels: Including a lag in panel data estimation significantly complicates the analysis. We may cover this topic if time remains at the end of the semester.

One of the themes of this course is that time conveys information that an econometrician may attempt to exploit. The method for exploiting this information in a panel that we have discussed is to include time fixed effects. This is far cruder than the methods discussed earlier when examining VARs, vector error correction models, and IRFs. We must decide in each case whether or not it is good enough.

If not, then we may wish to employ a panel VAR. These are demanding.