LECTURE NOTE 8: PANEL DATA

1 Introduction

The key ingredient to *panel data* is that observations are grouped in some important way. Often, we will follow a group of units (individuals, states, countries, etc.) over time, in which case we will write Y_{it} to denote the outcome of unit i in time period t. Such data are also called *longitudinal data*. In other instances, we will study groups of observations that have no ordering we consider relevant, for example siblings, classmates, and residents of the same neighborhood. One can think of possible orderings of these groups (birth order, class rank, etc.), but in many applications, it will be reasonable to ignore order. To distinguish the unordered groups from the ordered groups, we will write Y_{ij} to denote the outcome of (unordered) observation j from group i. The methods below apply to both types of panel data, but we will focus first on the unordered case because the algebra is somewhat simpler.

To clarify the discussion, we will frequently refer to two examples. For the unordered case, we will consider the determinants of earnings in a sample of sibling groups (which lay people also call "families"). For the ordered case, we will consider the relationship between traffic regulations (speed limits, seat belt laws, etc.) and motor vehicle fatalities in a sample of states over time.

2 Error Components Models

We begin by studying a large number of groups i, each including observations j. In the siblings example introduced above, i indexes families, and j indexes siblings within each family. Our basic regression model is:

$$Y_{ij} = \alpha + X_{ij}'\beta + Z_i'\gamma + U_{ij} \quad i = 1, \cdots, N \quad j = 1, \cdots, J \tag{1}$$

Y_{ij} is the outcome for the j^{th} observation from group i (e.g., log earnings in adulthood), X_{ij} is a vector of observation-specific characteristics (e.g., educational attainment), and Z_i is a vector of group-wide characteristics (e.g., parental education and wealth). Note that we have assumed a constant number of observations per group (siblings per family), J. If a panel satisfies this assumption, we call it a *balanced panel*; if it does not (so that instead of J we have J_i), we call it an *unbalanced panel*. One can apply the methods in this lecture to both balanced and unbalanced panels. However, in a balanced panel, the overall number of observations

is NJ, which is much more convenient than the corresponding number for an unbalanced panel. To simplify the algebra, we will assume a balanced panel in most of our discussion.

We might be tempted to run regression (1) using OLS, but the grouping of observations makes standard OLS inappropriate. To see the problem, it is useful to rewrite the residual U_{ij} as the sum of two components:

$$U_{ij} = \delta_i + \varepsilon_{ij} \tag{2}$$

where δ_i is the part of the residual that is the same for all observations within a group, and ε_{ij} is the part of the residual that is observation-specific. We assume that δ_i and ε_{ij} have mean zero and are uncorrelated with each other. In the siblings example, we would call δ_i a "family effect;" it would contain unobserved family-level variables such as shared ability, parental work ethic, parenting skills, and so on. Meanwhile, ε_{ij} would contain unobserved sibling-specific variables like a sibling's own IQ, own school experience, and so on.

We can rewrite equations (1) and (2) as follows:

$$Y_{ij} = \alpha + X_{ij}'\beta + Z_i'\gamma + \delta_i + \varepsilon_{ij} \tag{3}$$

We call equation (3) an *error components model*. The panel data methods we study below address two issues that arise in the estimation of the equation (3):

1. Error terms are positively correlated within each group: $cov(U_{ij}, U_{ik}) = V[\delta_i] > 0$ for $j \neq k$.
2. The δ_i component *may* be correlated with covariates X_{ij} and Z_j: $cov(\delta_i, X_{ij}) \neq 0$ or $cov(\delta_i, Z_i) \neq 0$.

Issue (1), which is *always* a concern in panel data, only affects the variance (standard errors) of our coefficient estimators. In the absence of a correlation between the error term and the covariates (i.e., when issue (2) is not a concern), OLS still delivers unbiased and consistent coefficient estimates, but they are inefficient, and the standard errors are incorrect. We have already learned one way to obtain correct standard errors: clustering. The clustered standard errors we studied in Lecture Note 4 allow for any within-group serial correlation, including the form that arises in issue (1). However, the OLS estimator with clustered standard errors is not efficient. We will briefly discuss an alternative method, *random effects estimation*, which is efficient. Issue (2) is a form of omitted variables bias: a correlation between the error term and the covariates. The panel structure of the data allows us to correct for this bias in a very sensible way, called *fixed effects estimation*.

3 Fixed Effects Estimation

The idea of fixed effects estimation is that we control for observation ij's membership in group i, so that our coefficient estimates are based only on *within-group* variation. We can do so in two ways. The first way,

often called the "brute force approach," involves two steps. First, generate $N-1$ dummy variables:

$$D_i = \begin{cases} 1 & \text{if observation is in group } i \\ 0 & \text{otherwise} \end{cases}$$

for $i = 2, \cdots, N$. Then control for the dummies directly:

$$Y_{ij} = \alpha + X_{ij}'\beta + \sum_{i=2}^{N} \lambda_i D_i + \varepsilon_{ij} \tag{4}$$

Note that once we control for D_i, we cannot estimate coefficients for Z_i because it is collinear with the full vector of D_i dummies. Thus, fixed effects estimation can only handle covariates that vary within group. We interpret β as the association between Y_{ij} and X_{ij}, holding fixed all variables that are shared within each group.

The "brute force" approach works well when we have relatively few groups, but when N is large, regression (4) becomes computationally burdensome. In that case, we can use the "finesse" method, which involves first demeaning Y_{ij} and X_{ij} within each group and then running a regression using the demeaned data. To understand this method, first collapse equation (3) to group-level means:

$$\bar{Y}_i = \alpha + \bar{X}_i'\beta + \bar{Z}_i'\gamma + \delta_i \tag{5}$$

Here, we have averaged out all of the within-group variation in our data. Equation (5) is known as the *between regression* because it uses only between-group variation. Now subtract equation (5) from equation (3):

$$Y_{ij} - \bar{Y}_i = \left(X_{ij} - \bar{X}_i\right)'\beta + \varepsilon_{ij} \tag{6}$$

This equation is known as the *within regression* because it uses only within-group variation. Both Z_i and δ_i have dropped out, so omitted variables bias from δ_i is no longer a concern. We can estimate the within regression by first estimating group-level means of Y_{ij} and X_{ij}, then calculating each observation's deviation from its group-level mean, $Y_{ij}^* = Y_{ij} - \bar{Y}_i$ and $X_{ij}^* = X_{ij} - \bar{X}_i$, and then running a regression of Y_{ij}^* on X_{ij}^*. This approach yields *identical* coefficient estimates to the "brute force" approach. However, the standard errors in the within regression need to be adjusted for the degrees of freedom used to compute $\bar{X}_i$.

In R, you can turn `groupvar` into a factor variable using `factor(groupvar)` and then apply the "brute force" approach:

- `feols(y ~ x + factor(groupvar), data = df)`

Or, even better, you can apply the "finesse" approach using the fixed effect syntax of `fixest`:

- `feols(y ~ x | groupvar, data = df)`

where `df` is the name of the dataframe. You can also apply the "finesse" approach using the `plm()` function from the `plm` package, setting `model = "within"`. In Stata, you can use `reg` with dummy variables, or you can use `areg` or `xtreg`.

4 Random Effects Estimation

When δ_i is uncorrelated with X_{ij} and Z_i, we can use both between and within variation to estimate $\hat{\beta}$. OLS estimation of equation (1) uses both sources of variation but is not not efficient. Random effects estimation attains efficiency by weighting the between (equation 5) and within (equation 6) components of the model to minimize the variance of $\hat{\beta}$. We do not have time to go over random effects estimation in this course, but you should know that it only seeks to improve efficiency relative to OLS. It does not "fix" any omitted variables bias. That is to say, for random effects just like for OLS, we would need to assume no group or individual omitted variables. If you read a research article that claims to address group-level omitted variable bias with a random effects model, you should be skeptical.

5 Time in Panel Data

Up to this point, we have primarily discussed observations that are grouped but unordered. But we will often encounter panel data that are both grouped and ordered, especially when we observe a single entity (person, state, country, etc.) over time. We typically use the subscripts i and t when describing this type of panel data, which is also called longitudinal data. All of the models described above also apply to longitudinal data. However, we may now wish to also include controls that vary over time but not across entities, as well as a time component in the error term:

$$Y_{it} = \alpha + X'_{it}\beta + Z'_i\gamma + W'_t\lambda + \delta_i + \tau_t + \varepsilon_{it} \tag{7}$$

In our state traffic laws example (where Y_{it} is motor vehicle fatalities), Z_i and δ_i vary across states but not over time, while W_t and τ_t vary over time but are shared across states. Z_i and δ_i might capture state differences in driving culture, while W_t and τ_t might capture changes in automotive technology or changes in national traffic laws.[1] In practice, we usually control for time variation by including time fixed effects: that is, including a separate dummy variable for each year (excluding the first as the base category). When we

[1] Z_i and W_t are measured, while δ_i and τ_t are not.

control for both state and time fixed effects, the estimating equation appears as:

$$Y_{it} = \alpha + X'_{it}\beta + \delta_i + \tau_t + \varepsilon_{it} \tag{8}$$

In the traffic laws example, the δ_i fixed effect controls for all time-invariant characteristics of state i, while the τ_t fixed effect controls for all national time trends that are shared by all states. Equation (8) is known as a *two-way fixed effects* regression. If X_{it} is a dummy for a seat-belt law, then β measures a state's change in fatalities that is associated with a change in seat-belt laws, net of national trends in fatalities over the same period. We are effectively using time trends in states that did not change their seat-belt laws as controls. Section 7 below will make this notion more explicit.

In R, you can transform the time variable to a factor variable using `factor(timevar)` and then apply the "brute force" approach:

- `feols(y ~ x + factor(groupvar) + factor(timevar), data = df)`

Or you can again apply the "finesse" approach using the very elegant fixed effect syntax of `fixest`:

- `feols(y ~ x | groupvar + timevar, data = df)`

where `df` is the name of the dataframe. Or you can use the `plm()` function from the `plm` package. Stata implementation uses the same commands from Section 3.

6 First Differencing

For longitudinal data, first difference estimation offers an alternative to fixed effects estimation. Consider a first-differenced version of equation (8):

$$\begin{aligned}
\Delta Y_{it} &= Y_{it} - Y_{i,t-1} \\
&= (\alpha + X'_{it}\beta + \delta_i + \tau_t + \varepsilon_{it}) - \left(\alpha + X'_{i,t-1}\beta + \delta_i + \tau_{t-1} + \varepsilon_{i,t-1}\right) \\
&= (X_{it} - X_{i,t-1})'\beta + (\tau_t - \tau_{t-1}) + (\varepsilon_{it} - \varepsilon_{it-1}) \\
&= \Delta X'_{it}\beta + \Delta\tau_t + \Delta\varepsilon_{it}
\end{aligned}$$

The δ_i fixed effect drops out of the regression, so we can also use first differencing to estimate β from within variation. To account for the $\Delta\tau_t$ component of the error term, we include dummies for each time period t. In the two-period case, the first difference and fixed effects estimators are identical. With more than two periods, the estimators yield different results, but both are consistent. Fixed effects estimation is more efficient when the ε_{it} error terms are i.i.d., while first difference estimation is more efficient when the ε_{it} error terms follow a random walk. In most situations, the reality lies somewhere between those two extremes, so neither method

is generally superior to the other. One practical consideration is that first differencing becomes difficult in unbalanced samples (i.e., when some time periods are missing for some states).

To carry out first difference estimation in R using `fixest`, you can use `d(var)` to take the first difference of a variable called `var` within state if you also specify the panel structure by adding `panel.id = ~state+year`:

- `feols(d(y) ~ d(x) | year, data = df, panel.id = ~state+year)`

You can also do first-difference estimation using the `plm()` function from the `plm` package. Stata has similar functionality to `fixest`. You declare the structure of your panel data using `xtset` and then apply the difference operator `D.var`.

7 Difference-in-Differences Estimation

Policy analysts very commonly use a related panel data research design called difference-in-differences. A new example will be useful. Suppose we are studying the effect of abortion bans on women's risk of dying in pregnancy. Some US states passed abortion bans when they were prohibited by the Supreme Court, but a recent Supreme Court decision allowed them to go into effect. Suppose we have a sample of many pregnant women (i) living in all 50 states (s) over several years (t) before and after the Supreme Court decision. The variable Y_{ist} equals 1 if woman i from state s in year t died in pregnancy, 0 if not.

Many policy analyses proceed with either a cross-sectional comparison or a time-series comparison. The cross-sectional approach would compare maternal mortality rates in ban states with maternal mortality rates in non-ban states after the Supreme Court decision:

$$\Delta_s = \bar{Y}_{BAN,POST} - \bar{Y}_{NO\ BAN,POST}$$

The time-series approach would look at ban states over time:

$$\Delta_t = \bar{Y}_{BAN,POST} - \bar{Y}_{BAN,PRE}$$

Both of these approaches have problems. The cross-sectional comparison is biased by the many other state-level differences in the determinants of maternal mortality, while the time-series comparison is biased by national trends, for example.

To eliminate both types of bias, we can estimate a difference-in-differences model:

$$\Delta\Delta = \left(\bar{Y}_{BAN,POST} - \bar{Y}_{BAN,PRE}\right) - \left(\bar{Y}_{NO\ BAN,POST} - \bar{Y}_{NO\ BAN,PRE}\right)$$

Just as in the first-differenced models in Section 6, this estimator eliminates time-invariant differences across

states, as well as time effects that are shared by all states. The key assumption is that *BAN* and *NO BAN* would be on parallel trends in the absence of the bans. Under this *parallel trends assumption*, $\bar{Y}_{NO\ BAN,POST} - \bar{Y}_{NO\ BAN,PRE}$ tells us what change ban states would have experienced in the absence of the bans. In more formal language, $\bar{Y}_{NO\ BAN,POST} - \bar{Y}_{NO\ BAN,PRE}$ is an estimate of the *counterfactual*.

We can use the a regression to estimate $\Delta\Delta$:

$$Y_{ist} = \beta_0 + \beta_1 POST_t + \beta_2 BAN_s + \beta_3 POST_t * BAN_s + U_{ist} \tag{9}$$

where $POST_t$ is a dummy for after the Supreme Court decision, BAN_s is a dummy for state s having an abortion ban on the books before the Supreme Court decision, and $POST_t * BAN_s$ is their interaction. To see how this regression relates to difference-in-differences estimation, we consider each of the means in the expression for $\Delta\Delta$ separately:

- $\bar{Y}_{NO\ BAN,PRE} = \beta_0$
- $\bar{Y}_{NO\ BAN,POST} = \beta_0 + \beta_1$
- $\bar{Y}_{BAN,PRES} = \beta_0 + \beta_2$
- $\bar{Y}_{BAN,POST} = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Now calculate:

$$\Delta\Delta = \left(\bar{Y}_{BAN,POST} - \bar{Y}_{BAN,PRE}\right) - \left(\bar{Y}_{NO\ BAN,POST} - \bar{Y}_{NO\ BAN,PRE}\right) = \beta_3$$

So the coefficient on the interaction term equals the difference-in-differences estimator.

A more general version of equation (9) is a two-way fixed effects regression:

$$Y_{ist} = \tau_t + \delta_s + \beta POLICY_{st} + \varepsilon_{ist} \tag{10}$$

where $POLICY_{st}$ equals one in state s only after the health policy has been implemented in that state. τ_t and δ_s are year and state fixed effects. β in equation (10) has the same interpretation as β_3 in equation (9), but we can now include multiple time periods.

We can also possibly use equation (10) to study states that implemented abortion bans at different times (a so-called *staggered policy rollout*), treating states that did not implement reform between any pair of years $t-1$ and t as a "control group" for those that did. Equation (10) has been used to study countless staggered rollouts. However, applied researchers are currently grappling with a caveat to this approach: sometimes already-treated states are *not* a valid "control group" for newly-treated states. For example, if the effects of the ban build over time, then a state that implemented a ban two years ago may still be experiencing policy

effects that push it off of pre-existing trends. Such a state would not be expected to be on parallel trends to a state that is currently implementing a ban. In contrast, untreated states are more likely to satisfy the parallel trends assumption. In staggered designs, two-way fixed effects assumes parallel trends for both untreated and already-treated states.

We conclude with a comment about standard errors. In the 1980s and 1990s, researchers often estimated equation (10) under the assumption that ε_{ist} is i.i.d. across observations. But it may not be i.i.d. for several reasons. First, individuals from the same state and year may have correlated errors, so that $cov(\varepsilon_{ist}, \varepsilon_{jst}) \neq 0$ for $i \neq j$. Second, because policy changes are very persistent, the error terms may be serially correlated over time within each state. First differencing offered one way to deal with this problem, but another is to cluster by state across all time periods (and across all individuals within each state). In R, `fixest` already does this. We may have millions of individual observations in our dataset, but we are assuming only 50 independent entities. Clustering works well if the data have at least 40 clusters but can be unreliable with a smaller number of clusters.