14.32 Econometric Data Science
Professor Anna Mikusheva
September, 2019

Lectures 17-18

Panel data

Panel data contains observations on multiple entities (individuals) observed for several periods in time. We will use two sets of indexes: $i = 1, ..., n$ stays for entity, $t = 1, ..., T$ stays for time. We have the following data:

$$\{(Y_{it}, X_{it}), i = 1, ..., n, t = 1, ..., T\} = \{(Y_{it}, X_{1,it}, ..., X_{k,it}), i = 1, ..., n, t = 1, ..., T\}.$$

We will work with balanced panels only, that is, the ones in which observations for all (i, t)-combinations are available. We will use panel data to control a special type of omitted variable bias.

Running example. The effect of alcohol taxes on traffic fatalities (through reducing drunk driving). Available data: on 48 US states for 7 years (1982-1988) on traffic fatalities (Y), alcohol taxes (X), demographics, etc.

A simple bivariate regression

$$Y = \alpha + \beta X + e$$

does not provide a valid estimate of the causal effect due to OVB. An example of omitted variable: cultural attitude towards drunk-driving or drinking culture:(1) keeping everything else fixed it is a determinant of fatalities, (2) those states have lower alcohol taxes.

Panel data allows us to correct for OVB due to variables that do not change over time.

Idea. Imagine that there is variable Z which is constant over time, the inclusion of which correct the bias. That is,

$$Y_{it} = \alpha + \beta X_{it} + \gamma Z_i + e_{it}$$

gives a valid estimate of β. Now assume that variable Z is not available. Apparently, the presence of a panel data structure allows us to resolve this problem. Let us write this model for two dates:

$$Y_{i1} = \alpha + \beta X_{i1} + \gamma Z_i + e_{i1},$$
$$Y_{iT} = \alpha + \beta X_{iT} + \gamma Z_i + e_{iT},$$

and take the difference:

$$Y_{iT} - Y_{i1} = \beta(X_{iT} - X_{i1}) + (e_{iT} - e_{i1}),$$

which does not include Z anymore. Thus, to estimate β consistently we can run a regression of $Y_{iT} - Y_{i1}$ on $X_{iT} - X_{i1}$.

Fixed effects

Imagine that the ideal regression one wants to run is

$$Y_{it} = \alpha + \beta X_{it} + \gamma Z_i + e_{it},$$

In the presence of unobserved Z_i this is equivalent to allowing different intercepts for different individuals: $\alpha_i = \alpha + \gamma Z_i$, and estimating model:

$$Y_{it} = \alpha_i + \beta X_{it} + e_{it}, \tag{1}$$

such a model is known as a 'fixed effect model', while individual intercepts α_i are called 'individual fixed effects'. Another way of writing the same model would be to introduce individual dummies D_j's – one for each individual $i = 1, .., n$:

$$D_{j,it} = \mathbb{I}\{j = i\},$$

Then the fixed-effect model is

$$Y = \alpha_1 D_1 + ... + \alpha_n D_n + \beta X + e.$$

Alternatively (remember multicollinearity!), we may keep an intercept and use $n - 1$ dummies:

$$Y = \alpha_0 + \alpha_1 D_1 + ... + \alpha_{n-1} D_{n-1} + \beta X + e.$$

There are two ways of estimating this model:

(1) Run the usual OLS on X and individual dummies – challenge: the matrix of regressors becomes too high-dimensional.

(2) De-meaning: Let us average equation (1) over time for each individual, that is, let us define $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ and other averages in a similar manner. Then:

$$\bar{Y}_i = \alpha + \beta \bar{X}_i + \bar{e}_i.$$

Now, let us subtract the averaged equation from equation (1):

$$Y_{it} - \bar{Y}_i = \beta(X_{it} - \bar{X}_i) + (e_{it} - \bar{e}_i).$$

Let us define de-meaned variables $\widetilde{Y}_{it} = Y_{it} - \bar{Y}_i$, and others in a similar manner. Then we have:

$$\widetilde{Y}_{it} = \beta \widetilde{X}_{it} + \widetilde{e}_{it}.$$

Thus, we may do the following : (1) de-mean all variables; (2) run OLS of $\widetilde{Y}_{it}$ on $\widetilde{X}_{it}$ without an intercept.

Claim. Running OLS with individual dummies and de-meaned regression give exactly the same value of the estimator for β.

Observation. If $T = 2$, then the differencing we did before: regressing $Y_{i2} - Y_{i1}$ on $X_{i2} - X_{i1}$ is equivalent to de-meaning.

Time fixed effect

Imagine that there are also some variables that may cause omitted variable bias, that are changing over time, but are constant over individuals. In our alcohol tax example, it may be national trends in car safety and trends towards stricter laws. Then the ideal regression may look like:

$$Y_{it} = \alpha + \beta X_{it} + \delta S_t + e_{it}.$$

If S_t is unobserved, estimation of such a model is equivalent to assuming time fixed effect $\mu_t = \alpha + \delta S_t$ or a different intercept for different time periods:

$$Y_{it} = \mu_t + \beta X_{it} + e_{it}.$$

One can introduce time dummies (one for each time period) $B_1, .., B_T$, and the model can be written as

$$Y = \mu_1 B_1 + ... + \mu_T B_T + \beta X_{it} + e_{it}.$$

Such a model can be estimated either by running OLS on a regression with time dummies, or by de-meaning by averaging over individuals.

Both time and individual fixed effects. By the logic above, we have a case for omitted variables of two types: ones that are constant over time, but different for different entities (like culture) and ones that are the same for all entities, but vary over time (like time trends). Thus, it makes sense to have both time fixed effects and individual fixed effects:

$$Y_{it} = \alpha_i + \mu_t + \beta X_{it} + e_{it}.$$

Notice that the levels of α_i's and μ_t's are not separately identifiable, and we typically do not have an interpretation for them. One can estimate the model by running a regression either including dummies for time/individuals or by de-meaning over time/individuals or combinations of them. Most commonly, we include time dummies but de-mean over time data for each individual. This is usually most useful, as we have data with many entities but that is quite short in terms of time dimension.

Assumptions for fixed effects

Assume that

$$Y_{it} = \alpha_i + \beta X_{it} + e_{it}$$

and that the following assumptions hold:

Assumption 1 $E(e_{it}|\alpha_i, X_{i1}, ..., X_{iT}) = 0$;

Assumption 2 $(X_{i1}, ..., X_{iT}, Y_{i1}, ..., Y_{iT})$ are i.i.d. draws from an unknown distribution;

Assumption 3 No outliers: 4-th moments of all regressors and errors are bounded;

Assumption 4 No perfect multicollinearity.

Statement. If Assumptions 1-4 hold, then the de-meaned OLS estimate for $\widehat{\beta}$ (OLS in the regression with individual dummies) is consistent and asymptotically gaussian.

Discussion of assumptions. Notice that Assumption 1 requires e_{it} to be uncorrelated with the WHOLE history of regressors for the i-the entity, not just X_{it}. This is hugely important, since we do de-meaning over time. This means that error e_{it} was unexpected given the past value of the regressors, but also does not produce any feedback (change in future X_{is}). For example, if in reaction to high traffic fatalities the local government is forced to increase its alcohol tax rate, the assumption will be violated. It is easier to satisfy an assumption for a shorter panel.

Notice that Assumption 2 requires i.i.d. sampling only over entities, but not for one entity over time. The latter would be too strong of an assumption to be useful in any real-life application.

Clustered Standard errors

One challenge with inferences for panel data is that error terms for the same entity (but different time periods) are correlated. For example, the error contains the unexplained part of traffic fatalities, such as road renovation or weather conditions; they tend to persist over several years, that would make e_{is} and e_{it} correlated. The usual heteroskedasticity-robust standard errors rely on an assumption that all errors (for all different combinations of (i, t)) are uncorrelated, and thus, may not work correctly if correlation within any entity is present.

The proper way of calculating standard errors is called 'clustered standard errors'- it allows for arbitrary error correlation within an entity, though assumes independence across entities (assumption 2). Clustered errors is one form of HAC (heteroskedasticity and auto-correlation robust) standard errors.

Idea. Assume for simplicity that we have one regressor (it all is easily generalizable to multiple regressors).

$$\sqrt{nT}(\widehat{\beta}-\beta)=\sqrt{nT}\left(\frac{\sum_{i=1}^n\sum_{t=1}^T\widetilde{X}_{it}\widetilde{Y}_{it}}{\sum_{i=1}^n\sum_{t=1}^T\widetilde{X}_{it}^2}-\beta\right)=\frac{\frac{1}{\sqrt{nT}}\sum_{i=1}^n\sum_{t=1}^T\widetilde{X}_{it}\widetilde{e}_{it}}{\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T\widetilde{X}_{it}^2}.$$

The derivation below assumes that T is small (relatively) and taken as fixed. In practice, almost all economic data sets are very short in time dimension. For the denominator we have:

$$\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T\widetilde{X}_{it}^2=\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{T}\sum_{t=1}^T\widetilde{X}_{it}^2\right)=\frac{1}{n}\sum_{i=1}^n\xi_i\to^p Q=E[\xi_i].$$

Here we use the Law of Large Numbers applied to i.i.d variables $\xi_i=\frac{1}{T}\sum_{t=1}^T\widetilde{X}_{it}^2$. Similarly,

$$\frac{1}{\sqrt{nT}}\sum_{i=1}^n\sum_{t=1}^T\widetilde{X}_{it}\widetilde{e}_{it}=\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T\widetilde{X}_{it}\widetilde{e}_{it}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\eta_i\Rightarrow N(0,Var(\eta_i)),$$

where we used the Central Limit Theorem for i.i.d. variables $\eta_i=\frac{1}{\sqrt{T}}\sum_{t=1}^T\widetilde{X}_{it}\widetilde{e}_{it}$, which is mean-zero due to Assumption 1.

A proper way to estimate $Var(\eta_i)$ is to calculate the sample variance for produced random variables η_i.