Introduction to Econometrics: Recitation 6

Seung-hun Lee

November 15th, 2021

1 Panel Regression

1.1 Motivation for Using Panel Data

We now get to analyze a different type of dataset. We are moving into the **panel data**, where we observe multiple individuals for multiple periods of time. In terms of notation, we write

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + ... + \beta_k X_{k,it} + u_{it}$$

where $i = 1, 2, ..., N$ indicates individuals and $t = 1, 2, ..., T$ indicates time periods. For some panel datasets, there are T datasets for each of the N individuals included in the data. This type of panel data is said to be a **balanced panel data**. However, most panel data available to public has cases where there are $t \leq T$ datapoints for some of the N individuals. This type of dataset is an **unbalanced panel data**.

There are many reasons for using a panel data. One is simply that using a panel data allows us to use more datasets. As we have learned from the previous lectures, our estimates get easier and more accurate as we have more data. However, a crucial advantage of using a panel dataset comes from the fact that it can allow us to control for a particular type of omitted variable bias. Specifically, it can allow us to control for **unobserved heterogeneity** that are either 1) different accross N entities but always remain same for T periods in a given entity or 2) different accross T times periods but remains the same for all N entities in a particular time period or 3) both of 1) and 2). The first of this is called a **cross section (entity) fixed effect**. The second one is **time fixed effects**. Case 3) is what we call **two-way fixed effects**.

For curious minds: Can the unobservable heterogeneity be unrelated to the independent variables? In theory, yes. This type of unobserved heterogeneity is commonly referred to as **random effects**. In reality, not many data satisfy this. So we will delay the discussion of this topic to an advanced course

To see why that is the case, suppose that $T = 2$ and we are interested in the relationship between vehicle related fatality rate (deaths per 10,000 people) and the beer tax. Suppose that we get these result for the two years

$$\begin{array}{cc} \hat{Y}_{i1} = 2.01 & +0.15X_{i1} \\ (0.15) & (0.20) \\ \hat{Y}_{i2} = 1.86 & +0.44X_{i2} \\ (0.11) & (0.20) \end{array}$$

In the first year, the coefficient on X_i is not significant at all, whereas the same coefficient for year 2 is very significant. In such case, one might suspect that there is an omitted variable bias that affects these coefficients. For instance, if i denotes states, one might guess that some type of omitted variable that are specific to the states could be affecting these results - like strictness of DUI laws in some state i. If it is the case that the strictness of DUI laws in a given state did not change for the two time periods, then these effects qualify as a state fixed effect.

To see how such omitted variable bias can be controlled for, let Z_i denote the strictness of state laws on DUI. This is not exactly measurable. However, if there is no significant change in state law for the two periods, Z_i can be considered unchanging. Now write

$$\begin{aligned} Y_{i1} &= \beta_0 + \beta_1 X_{i1} + \beta_2 Z_i + u_{i1} \\ Y_{i2} &= \beta_0 + \beta_1 X_{i2} + \beta_2 Z_i + u_{i2} \end{aligned}$$

Subtract the second equation from the first to get

$$(Y_{i2} - Y_{i1}) = \beta_1(X_{i2} - X_{i1}) + \beta_2(Z_i - Z_i) + u_{i2} - u_{i1}$$

With Z_i being the same for all periods, the above equation is reduced to

$$(Y_{i2} - Y_{i1}) = \beta_1(X_{i2} - X_{i1}) + (u_{i2} - u_{i1})$$

The Z_i variable has no role in this equation. Effectively, we did control for state fixed effect that are constant across time. If we estimate this particular β_1, we can obtain much more accurate estimates of the effect of beer tax on fatality rate. The β_1 in this case can be interpreted as how much the change in $(X_{i2} - X_{i1})$ changes $(Y_{i2} - Y_{i1})$.

Let's use a numerical example, Suppose that beer tax is \$10 in year 1, so $X_{i1} = 10$. Suppose that state i raises its bear tax by 1 dollars, therefore $X_{i2} - X_{i1} = 1$. If β_1 is estimated as -1.1, then we can write the predicted change in Y (which would be a predicted change in death per 10,000 people) as

$$\widehat{Y_{i2} - Y_{i1}} = -1.1(X_{i2} - X_{i1})$$

If you put $X_{i2} - X_{i1} = 1$, Then $\widehat{Y_{i2} - Y_{i1}} = -1.1$, or that fatality per 10,000 reduces by 1.1. Also

notice that this is starkly different result compared to the estimates we got when we regressed year by year (coefficient in front of X was positive, as opposed to a negative coefficient for change in X we used for the second approach).

1.2 Methodology

We are going to limit our discussion to a cross-sectional fixed effects, as all logic would hold by changing cross sectional index i to time index t. For each of these fixed effects, there are two ways of estimating the data when $T \geq 3$. One of them is to include $N-1$ individual dummy variables (or $T-1$ time period dummy variables), sometimes referred to as **least square dummy variable method**. The other is to subtract from the original equation the "demeaned" equation, which is referred to as **within estimation**.

For both approaches, the following framework will work. Assume that the true relation between Y and X variable is

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it} \tag{1}$$

where Z_i is the cross section fixed effect. For $i = 1, ..., N$, the value of $\beta_2 Z_i$ is different. If we were to express this equation on an (X, Y)-plane, we would end up with a different intercept for each i. So for simplification, define $\alpha_i = \beta_0 + \beta_1 Z_i$. Then the above equation can be written as

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \tag{2}$$

in equation (2), the α_i term can be thought of as an effect of being an entity i. Since this is correlated with X_{it}, we deal with this in two ways

1.2.1 Least Square Dummy Variable Method

For this, we are going to make changes to equation (1). Define a new variable D_{ki} as follows

$$D_{ki} = \begin{cases} 1 & \text{If } i = k \\ 0 & \text{Otherwise} \end{cases}, \; k \in \{1, 2, ..., N\}$$

This variable takes 1 if i is the kth entity. We are defining this for all N entities. Since we are going to include β_0, a common intercept, in our regression we need to remove one of the N D_{ki} variables out to avoid dummy variable trap. Let's remove D_{1i} from the equation and include all others, then equation (1) becomes

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 D_{2i} + ... + \delta_N D_{Ni} + u_{it} \tag{LSDV}$$

To see what this equation gives, we analyze what the intercept will be for each of the N entities. This can be written out as

$$\begin{aligned} i = 1 &\implies (1) : \beta_0 + \beta_2 Z_1 \quad (2) : \alpha_1 \; (3) : \beta_0 \\ i = 2 &\implies (1) : \beta_0 + \beta_2 Z_2 \quad (2) : \alpha_2 \; (3) : \beta_0 + \delta_2 \\ &\quad \ldots\ldots \qquad\qquad \ldots\ldots \\ i = N &\implies (1) : \beta_0 + \beta_2 Z_N \quad (2) : \alpha_N \; (3) : \beta_0 + \delta_N \end{aligned}$$

In all of these cases, slope coefficient in front of X_{it} remains the same. In addition, we control for unobserved cross section fixed effect by allowing the intercept to differ by each i.

One clear disadvantage of this method is the computational burden when N or T is large. If N is extremely large, this cannot be solved on pen(cil) and paper. It may even take computer quite a lot of time to process the results. The computational burden cam be eased with the following method.

1.2.2 Within Transformation Method

Go back to equation (2). Since we are concerned about cross section fixed effects, we define a new term, $\bar{X}_i, \bar{Y}_i$ as sample mean of X_{it}, Y_{it} for some given i over all possible t's. This can be calculated as

$$\bar{X}_i = \frac{1}{T} \sum_{t=1}^{T} X_{it}$$

Consequently, $\bar{Y}_i$ can be written as

$$\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it} = \frac{1}{T} \sum_{t=1}^{T} (\beta_1 X_{it} + \alpha_i + u_{it}) = \beta_1 \bar{X}_i + \alpha_i + \bar{u}_i$$

where $\bar{u}_i$ is defined in a similar manner to $\bar{X}_i, \bar{Y}_i$. Then, we subtract Y_{it} by $\bar{Y}_i$ to get

$$\begin{aligned} & Y_{it} - \bar{Y}_i = \beta_1 (X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i) \\ \implies & \tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \end{aligned}$$

where $\tilde{X}_{it} = (X_{it} - \bar{X}_i)$. This controls the fixed effect term by effectively getting rid of them through this transformation process (also known as within transformation). To get the β_1 estimate, we solve the minimization problem similar to what we have went through in finding the original OLS estimator. This gets us

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{X}_{it} \tilde{Y}_{it}}{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{X}_{it}^2}$$

Within estimator: Again, the minimization problem is to find the $\hat{\beta}_1$ that minimizes the sum of the squared residual terms. This is written as

$$\min_{\hat{\beta}_1} \sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{Y}_{it} - \hat{\beta}_1 \tilde{X}_{it})^2$$

Differentiate this term with respect to $\hat{\beta}_1$ to get this first order condition

$$-2 \sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{Y}_{it} - \hat{\beta}_1 \tilde{X}_{it}) \tilde{X}_{it} = 0$$

Solving this, we can get $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{X}_{it} \tilde{Y}_{it}}{\sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{X}_{it}^2}$.

For time fixed effects, all of our logic remains similar. The exception is that we now need to use $\bar{Y}_t$, defined as

$$\bar{Y}_t = \frac{1}{N} \sum_{i=1}^{N} (Y_{it}) = \beta_1 \bar{X}_t + \alpha_t + \bar{u}_t$$

1.2.3 Time and Entity Fixed Effects

This is sometimes called a **two-way fixed effect models**. This is written by

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

As with the previous fixed effect models, there are two ways to deal with this

- **Least square dummy variable**: Now, we include $N - 1$ and $T - 1$ binary independent variables, along with a common intercept β_0 to get

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 T_{2t} + ... + \gamma_T T_{Tt} + \delta_2 D_{2i} + ... + \delta_N D_{Ni} + u_{it}$$

- **Within estimator**: This is a bit complicated, but possible. When both time and entity fixed effects are present, the way of demeaning the data is different. We first demand by time and entity, and then add back the overall sample mean back. In other words, the demeaning process is

$$Y_{it} - \bar{Y}_i - \bar{Y}_t + \bar{Y}$$

  where $\bar{Y} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it}$ and

  - $\bar{Y}_i = \beta_1 \bar{X}_i + \alpha_i + \frac{1}{T} \sum_{t=1}^{T} \lambda_t + \bar{u}_i$

- $\bar{Y}_t = \beta_1 \bar{X}_t + \frac{1}{N}\sum_{i=1}^{N} \alpha_i + \lambda_t + \bar{u}_t$
- $\bar{Y} = \beta_1 \bar{X} + \frac{1}{N}\sum_{i=1}^{N} \alpha_i + \frac{1}{T}\sum_{t=1}^{T} \lambda_t + \bar{u}$

Then, we get

$$(Y_{it} - \bar{Y}_i - \bar{Y}_t + \bar{Y}) = \beta_1(X_{it} - \bar{X}_i - \bar{X}_t + \bar{X}) + (u_{it} - \bar{u}_i - \bar{u}_t + \bar{u})$$

We conduct the OLS estimation here to get the estimate for β_1.

1.2.4 Least Square Assumption for Panel Data

The following assumptions are extensions from what we had in the cross sectional data. However, there are some key differences that are observed in the panel data setting. The four main assumptions are

P1 : $E[u_{it}|X_{i1}, .., X_{iT}, \alpha_i] = 0$. It means that the conditional mean of the u_{it} term does not depend on any of the X_{it} values for entity i, whether in the future or in the past. This rules out omitted variable bias. Some books refer to this as strict exogeneity.

P2 : $(X_{i1}, .., X_{iT}, u_{i1}, ...u_{iT})$ is IID across $i = 1, .., n$. In other words, variables for one entity are distributed identically to but independent of other entities. However, **this does not rule out the correlation between u_{it}, u_{ij} within entity i for different j and t**. Therefore, within an entity, there could be an autocorrelation (or serial correlation).

P3 : (X_{it}, u_{it}) have nonzero finite fourth moments. In other words, outliers are very unlikely. We need this condition to derive the asymptotic distribution of the estimators in the panel regression. (Not a scope of this class)

P4 : There is no perfect multicollinearity

Because of **P2**, the standard errors should be calculated while taking into account the possibility of autocorrelation within an entity. This is where **clustered standard error** comes in. This type of standard errors allow for heteroskedasticity and autocorrelation within an entity. However, it assumes that regression errors are independent across different entities. In practice, clustered standard errors are larger than regular standard errors (but not always). As in the case of heteroskedasticity, we should care about clustering as using "wrong" standard errors would lead us to making a wrong judgement about the hypothesis we set on some variables.

2 Binary Dependent Variables

2.1 Linear Probability Models

We now turn to the case where the dependent variable y takes either 0 or 1. This type of regression can be used to study how independent variable(s) X_i is(are) correlated to yes/no questions in the survey. As always, assume a following regression equation

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where Y_i is a binary variable. Unlike previous regressions, the complication arises when we attempt to interpret the equation. Especially, what does β_1 now mean? To study this question, we first look into the expected value of Y_i conditional on X_i, $E[Y_i|X_i]$. The conditional mean of Y_i is, by definition

$$E[Y_i|X_i] = 0 \times \Pr(Y_i = 0|X_i) + 1 \times \Pr(Y_i = 1|X_i)$$

In the context of this regression equation, we can obtain the conditional mean of Y_i as follows

$$\begin{aligned} E[Y_i|X_i] &= E[\beta_0 + \beta_1 X_i + u_i|X_i] \\ &= \beta_0 + \beta_1 X_i + E[u_i|X_i] \\ (\because E[u_i|X_i] = 0) &= \beta_0 + \beta_1 X_i \end{aligned}$$

Therefore, we just established that $E[Y_i|X_i]$ in this context is the probability of $Y_i = 1$ given X_i.

Now we move back to β_1, notice that $\beta_1 = \frac{\Delta Y_i}{\Delta X_i}$ and $\Delta Y_i =$ Change in $\Pr(Y_i = 1|X_i)$ with respect to change in X_i , or

$$\Delta Y_i = \Pr(Y_i = 1|X_i = x + \Delta X_i) - \Pr(Y_i = 1|X_i = x)$$

and when we calculate $\Pr(Y_i = 1|X_i = x + \Delta X_i) - \Pr(Y_i = 1|X_i = x) = E[Y_i|X_i = x + \Delta X_i] - E[Y_i|X_i = x]$, we get

$$\beta_0 + \beta_1(x + \Delta X_i) - \beta_0 + \beta_1(x) = \beta_1 \Delta X_i$$

So we get $\Delta Y_i = \beta_1 \Delta X_i \iff \beta_1 = \frac{\Delta Y_i}{\Delta X_i}$. Therefore, β_1 now measures how much the predicted probability of $Y_i = 1$ changes with respect to X_i

The **linear probability model** is the estimation in which you run an OLS on the type of regression equation where Y_i is a binary dependent variable. The advantage is that it is simple - there is no difference in terms of methods between this and the OLS methods we have learned so far. However, there are some critical disadvantages to this model. One is

that by setting the regression model as above, we are assuming that the change of predicted probability of $Y_i = 1$ is constant for all values of X_i. But more critically, it is possible that because of the way functional form is specified, the predicted probability $\hat{y}$ may be greater than 1 or strictly less than 0. Given that probability is defined to be in between 0 and 1, this could be a preposterous result. In addition, the distribution of the error term is no longer normal distribution. This could affect the asymptotic (large sample) properties of the OLS estimators.

2.2 Logit and Probit Regressions

Since linear probability models can exceed 1 or fall below 0, we now use a class of **sigmoid estimators**. These estimators are bounded between 0 and 1. Therefore, using these estimators will prevent the predicted probability of $Y_i = 1$ falling out of $[0, 1]$ range.

One of such estimator is **logit regression**. For notational convenience, I write $Z_i = \beta_0 + \beta_1 X_i$. Logit regression assumes that the cumulative probability of Z_i, which is $\Pr(Y_i = 1|X_i)$ is distributed as

$$\Pr(Y_i = 1|X_i) = F(Z_i) = \frac{1}{1 + e^{-Z_i}}$$

In this setup, when $Z_i \to \infty, e^{-Z_i} \to 0$ Therefore, $F(Z_i) \to 1$. Likewise, taking $Z_i \to -\infty$ results in $F(Z_i) \to 0$.

To see the role of the independent variable X_i, we should note that changes in X_i leads to changes in Z_i, since $Z_i = \beta_0 + \beta_1 X_i$. The change in Z_i should also impact $F(Z_i)$. Borrowing the logic from the chain rule, we get

$$\frac{\partial F}{\partial X_i} = \frac{\partial F}{\partial Z_i} \frac{\partial Z_i}{\partial X_i}$$

where $\frac{\partial Z_i}{\partial X_i} = \beta_1$. This says that β_1 alone does not explain how changes in X_i alters probability of $Y_i = 1$ given X_i. Therefore, the coefficient value of β_1 does not mean that much in logit regression (similar for probit regressions). However, Note that $\frac{\partial F}{\partial Z_i}$ is calculated as

$$\frac{\partial F}{\partial Z_i} = \frac{e^{-\beta_0 - \beta_1 X_i}}{(1 + e^{-\beta_0 - \beta_1 X_i})^2}$$

Therefore the sign of β_1 still matters. In fact, the interpretation of logit coefficients (and probit) regression matters up to the sign of the coefficients in general.

Probit regression is largely similar with the logit regression, except now that the cumulative probabiltiy of Z_i is assumed to be a standard normal function. Specifically,

$$F(Z_i) = \Phi(Z_i) = \Phi(\beta_0 + \beta_1 X_i)$$

where $\Phi(v)$ means the cumulative normal function $\Pr(Z \leq v)$. Again, the value of β_1 coefficient does not mean as much as the sign. By taking the similar approach with the logit regression, we get

$$\frac{\partial F}{\partial X_i} = \frac{\partial F}{\partial Z_i}\frac{\partial Z_i}{\partial X_i}$$

where $\frac{\partial F}{\partial Z_i}$ is the pdf of a standard normal distribution.

In practice, the two regressions are run together to check whether the signs of the coefficients are the same. There are not many differences between the two.

2.3 Maximum Likelihood Estimation Method

Notice that both logit and probit regressions are nonlinear in the sense that the β_0, β_1 parameters are no longer in linear relationship with the $X_i's$ and subsequently Y_i's. One of the assumptions used in using OLS is that the linear regression assumption. This is no longer a valid option anymore, which requires a different approach. This is where **maximum likelihood estimation** comes in. To understand the maximum likelihood estimators, you must understand what the likelihood function is. A **likelihood function** is the conditional density of $Y_1, ..., Y_n$ given $X_1, ..., X_n$ that is treated as the function of the unknown parameters (β_0, β_1 in our case). In other words, since we have the observations for Y_i's and X_i's, but do not know the values of the parameter β_i's, what we are trying to do here is to find the values of β_i's that best matches the values of X_i's and Y_i's. As a result, the maximum likelihood estimators is the value of β_i's that best describes the data and maximizes the value of the likelihood function

To nail this home, let's not worry about regression equation for the moment and consider a single variable - Y_i. For example, Let's assume that Y_i's are IID normal with mean μ and standard error σ, both of which are unknown. The joint probability of Y_i's are (our likelihood function)

$$\begin{aligned}
\Pr(Y_1 = y_1, ..., Y_n = y_n|\mu, \sigma) &= \Pr(Y_1 = y_1|\mu, \sigma) \times .. \times \Pr(Y_n = y_n|\mu, \sigma) \\
&= \prod_{i=1}^{n} f(y_i|\mu, \sigma) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i-\mu)^2}{2\sigma^2}} \\
&= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n} \frac{(Y_i-\mu)^2}{2\sigma^2}}
\end{aligned}$$

A convenient way to solve this class of problem is to use the *log-likelihod function*. Taking logs

to above equation gets us

$$-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \sum_{i=1}^{n}\frac{(Y_i-\mu)^2}{2\sigma^2} \qquad (*)$$

To find the maximum likelihood estimator of μ, we differentiate the above with respect to μ. This gets us

$$\begin{aligned} 2\sum_{i=1}^{n}\frac{(Y_i-\mu)}{2\sigma^2} = 0 &\implies \sum_{i=1}^{n}\frac{(Y_i-\mu)}{\sigma^2} = 0 \\ (\because \sigma, \text{ though unknown, will be a constant }) &\implies \sum_{i=1}^{n}(Y_i-\mu) = 0 \\ &\implies n\mu = \sum_{i=1}^{n} Y_i \implies \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} Y_i \end{aligned}$$

We can do the similar for σ^2 by differentiating $(*)$ with respect to σ^2. This leads to

$$\begin{aligned} -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}\frac{2(Y_i-\mu)^2}{(2\sigma^2)^2} = 0 &\implies \frac{n}{2\sigma^2} = \sum_{i=1}^{n}\frac{(Y_i-\mu)^2}{2\sigma^4} \\ &\implies n = \sum_{i=1}^{n}\frac{(Y_i-\mu)^2}{\sigma^2} \\ &\implies n = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2 \end{aligned}$$

By imputing μ_{MLE} in place of μ, we get

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_{MLE})^2$$