Panel Data: Lecture 1

April 18, 2024

Recap

- Linear Regression is the best linear approximation of the CEF.

Recap

- Linear Regression is the best linear approximation of the CEF.
- CEF is the best predictor of Y. Thus, linear regression is the best linear predictor of Y

Recap

- Linear Regression is the best linear approximation of the CEF.
- CEF is the best predictor of Y. Thus, linear regression is the best linear predictor of Y
- Linear regression can capture non-linearities by controling for higher order polynomials of the explanatory variables. The important thing is that it is linear in parameters.

Recap

- Linear Regression is the best linear approximation of the CEF.
- CEF is the best predictor of Y. Thus, linear regression is the best linear predictor of Y
- Linear regression can capture non-linearities by controling for higher order polynomials of the explanatory variables. The important thing is that it is linear in parameters.
- The goal of econometrics is to find casual relationships. -¿ We want to compare actual outcomes to potential outcomes. E.g. of what would have happened if someone without college degree had a college degree?

Recap

- Linear Regression is the best linear approximation of the CEF.
- CEF is the best predictor of Y. Thus, linear regression is the best linear predictor of Y
- Linear regression can capture non-linearities by controling for higher order polynomials of the explanatory variables. The important thing is that it is linear in parameters.
- The goal of econometrics is to find casual relationships. -¿ We want to compare actual outcomes to potential outcomes. E.g. of what would have happened if someone without college degree had a college degree?
- To make casual inference, we need to get rid of endogenous (bad) variation in the explanatory variables.

Omitted Time Constant Variables

Suppose a regression

$$ln(wage_{it}) = \alpha + \rho Union_{it} + \gamma A_i + \beta X_{it} + e_{it} \quad (1)$$

- Union is a dummy variable equal to 1 if a worker belongs to a labour union and 0 otherwise
- A_i is a set of unobserved variables that do not change over time. (Example?)
- X_{it} is a set of observed variables that vary across individual and across time. (Example?)
- $ln(wage_i t)$ is natural logarithm of observed wage

However A_{it} is unobserved so instead we can only estimate

$$ln(wage_{it}) = \alpha + \rho Union_{it} + \beta X_{it} + u_{it} \quad (2)$$

What are the consequences?

Omitted Variable Bias

First make some adjustments of the matrices. Without loss of generality:

- Define $W_{it} = \begin{bmatrix} Union_{it} & X_{it} \end{bmatrix}$
- Gather all observations from W_{it} and $ln(wage_{it})$ into large matrices W and $ln(WAGE)$
- Rewrite equation (2)

$$ln(WAGE) = W\Theta + u \tag{3}$$

Where $\theta = \begin{bmatrix} \alpha & \rho & \beta \end{bmatrix}^T$

Omitted Variable Bias contd...

Now start with OLS formula

$$\begin{aligned}
\hat{\Theta} &= (W^T W)^{-1} W^T ln(WAGE) \\
&= (W^T W)^{-1} W^T (W\Theta + u) \\
&= (W^T W)^{-1} W^T (W\Theta + A\Gamma + e) \\
&= (W^T W)^{-1} W^T W\Theta + (W^T W)^{-1} W^T A\Gamma + (W^T W)^{-1} W^T e \\
&= \Theta + (W^T W)^{-1} W^T A\Gamma \\
&= \Theta + \Delta\Gamma
\end{aligned}$$

The last line implies that the estimates of the coefficients from the short regression (2) are equal to the estimates of the long regression (1) plus the effect of the omitted variables A times the effect of the omitted variables A on the included variables W.
Thus, if (1) is the true causal model. Then estimating (2) will lead to omitted variable bias. The direction of the bias depends on the signs of Γ and Δ.

Fixed Effects Model

Panel data allow us discard the variation in W that is due to the omitted variables A. To see this consider again equation (1).

$$ln(wage_{it}) = \alpha + \rho Union_{it} + \gamma A_i + \beta X_{it} + e_{it}$$

Let $\alpha_i \equiv \alpha + \gamma A_i$, Then

$$ln(wage_{it}) = \alpha_i + \rho Union_{it} + \beta X_{it} + e_{it} \quad (4)$$

- Note that this model has more parameters to estimate than there are number of individuals (N).
- Fortunately we do not need consistent estimates of α_i to obtain consistent estimates of ρ. We just need to "kill" the variation in Union and X that is related to the fixed effects α_i.

Fixed Effects Model: Deviations from the Mean

We can exploit panel data structure to get rid of the individual fixed effects α_i. Firstly, for every individual i, calculate averages of the variables in equation (4).

$$\overline{ln(wage_i)} = \alpha_i + \rho\overline{Union_i} + \beta\overline{X_i} + \overline{e_i} \tag{5}$$

where every variable V in the model (5) is calculated in the following way:
$\overline{V_i} = \frac{1}{T}\sum_t V_{i,t}$
Next subtract (5) from (4)

$$ln(wage_{it}) - \overline{ln(wage_i)} = \rho Union_{it} - \rho\overline{Union_i} + \beta X_{it} - \beta\overline{X_i} + e_{it} - \overline{e_i} \tag{6}$$

Equation (6) is the Fixed Effects model. It is sometimes called the Within estimator because it uses only the within unit (e.g. individual) variation.

Within Estimator: Caveat

- The within estimator is useful when we want to control for fixed effects. For example unobserved time constant variables α_i. But there is a caveat. Which?

Within Estimator: Caveat

- The within estimator is useful when we want to control for fixed effects. For example unobserved time constant variables α_i. But there is a caveat. Which?
- Within estimator can only estimate effects of variables that vary on both dimensions (e.g. i and t).
- Pooled OLS (POLS) uses all the variation (between and within) so it is able to estimate all variables but there is a risk of omitted variable bias.

Within Estimator or POLS?

Formalize the conditions: Consider the equation (4)

$$ln(wage_{it}) = \alpha_i + \rho Union_{it} + \beta X_{it} + e_{it}$$

Suppose we run POLS

$$ln(wage_{it}) = \alpha + \rho Union_{it} + \beta X_{it} + v_{it}$$

Suppose we do not observe some variables in α_i. The POLS will give us consistent estimates of ρ if the following holds:

$$cov(Union, v_i) = 0$$

Thus,

$$cov(Union, e_i) = 0$$

and

$$cov(Union, \alpha_i) = 0$$

Example

Suppose you have data on hourly wages of male workers in the U.S. Each of these men continuously worked from 1980 to 1987. You observe the following variables: education, experience, race, and whether a worker is a member of a working union or not. Suppose we came up with the following model.

$$ln(wage_{i,t}) = \beta_0 + \beta_1 educ_{i,t} + \beta_2 exper_{i,t} + \beta_3 exper_{i,t}^2 + \beta_4 black_i + \beta_5 hispan_i + \beta_6 union_{i,t} + \sum_{j=1981}^{1987} \gamma_j D_j + \epsilon_{i,t}$$

a. Suppose we decide to estimate the equation using Pooled OLS. What assumption on the error term should we impose to obtain consistent estimates?

b. Suppose we run FE regression. What is the underlying assumption on the error term now? Can we identify all the coefficients? Explain in details.

c. Suppose the union premium estimated by FE is by 10 % lower than the OLS estimate. What does this suggest about the correlation between union and the unobserved effect?

Example contd.

a Consistent estimate requires the $cov(x_{i,t}, \epsilon_{i,t}) = 0$ condition to hold. In panel data the error term $\epsilon_{i,t}$ is usually $\epsilon_{i,t} = \alpha_i + u_{i,t}$. Hence the consistency requires the following to hold:

$$cov(x_{i,t}, \alpha_i) = 0$$

and

$$cov(x_{i,t}, u_{i,t}) = 0$$

b The FE model relaxes the $cov(x_{i,t}, \alpha_i) = 0$ condition, but still requires the $cov(x_{i,t}, u_{i,t}) = 0$ condition to hold.
FE are not able to identify the parameters of the time invariant variables. Hence $\beta_0, \beta_4, and \beta_5$ are not identified.

c $\beta_5^{FE} < \beta_5^{OLS}$ suggests that the $cov(union_{i,t}, \alpha_i) > 0$

Estimates of the Individual Effects α_i

- To estimate consistently ρ in equation (4), we just need large number of individuals (N).
- However, to consistently estimate α_i's. We need large number of time periods T.

Proof of α_i Inconsistency With Fixed Time Periods

Suppose a model $y_i = x_i\beta + \alpha_i + u_i$, show that the estimate of α_i is inconsistent.

First, notice that an estimator is consistent if in the limit it approaches the true coefficient. Write this condition in the mean squared error sense:

$$MSE = E_\theta[(h(X) - \theta)^2] = V_\theta(h(x)) + (E[h(X)] - \theta)^2 = 0$$

Where $h(X)$ is estimator and θ is the true population coefficient.
Now write the equation for estimate of α_i

$$\hat{\alpha}_i = \overline{y_i} - \overline{x_i}'\hat{\beta}$$

$$Var(\hat{\alpha}_i) = \frac{1}{T}\sigma_u^2 + \overline{x_i}' Var(\hat{\beta})\overline{x_i}$$

If $\hat{\beta}$ is consistent, the second term of the variance equation goes to 0 as N goes to infinity. However the first term does not vanish unless T goes to infinity as well.

Excercise

Use the data in ATTEND.dta to answer this question.

a. To determine the effects of attending lecture on final exam performance, estimate a model relating *stndfnl* (the standardized final exam score) to *atndrte* (the percent of lectures attended). Include the binary variables frosh and soph as explanatory variables. Interpret the coefficient on *atndrte*, and discuss its significance.

b. How confident are you that the OLS estimates from part (a) are estimating the causal effect of attendance? Explain.

c. You are worried that you omitted student ability. Comment on the direction of the bias. Be sure to state all necessary assumptions and justify your reasoning on the sign of the coefficients necessary to quantify the bias.

d. As proxy variables for student ability, add to the regression *priGPA* (prior cumu- lative GPA) and *ACT* (achievement test score). Now what is the effect of *atndrte*? Discuss how the effect differs from that in part (a)?

e. Visualize the results from (a) and (d) in a graph using ggplot2.

Excercise contd.

Until next week, Answer upload a document with all the questions answered and upload a working R script.