MSA 8200 Predictive Analytics
Week8: Estimation Bias and Omitted Variable

Yichen Cheng

Spring 2020

Outline

- What is estimation bias?
- The omitted variable problem.
- Panel Data
- Random effect model vs fixed effect model.
- Difference in differences model

Motivating examples – Basketball +/- values

Evaluate the performance of a player:

- Compares how a team scores with a player on/off the court.
- Measures the overall performance of a player.

Is it a good measure?

Motivating examples - Airbnb

Evaluate Airbnb's new feature:

- A deep learning algorithm to improve photo's quality.
- The hosts can choose whether to use this feature.
- An analyst randomly selected 10,000 hosts and collect their information:

$$earn = \beta_1 + \beta_2(numbeds) + \beta_3(prog) + u \tag{1}$$

numbeds: total number of bedrooms of each property.

prog: binary indicator variable, 1 if the host used the new feature.

Estimation Bias

Consider a linear regression case:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

An estimate ($\hat{\beta}_1$) is said to be unbiased if:

$$E(\hat{\beta}_1) = \beta_1$$

i.e. the expected value of the estimate is equal to the truth.

An estimate ($\hat{\beta}_1$) is said to be consistent if:

$$E(\hat{\beta}_1) \rightarrow \beta_1 asn \rightarrow \infty$$

i.e. the expected value of the estimate approaches the truth for large sample.

Estimation Bias

Condition for the OLS estimator to be unbiased:

$$E(\epsilon|x) = 0 \tag{3}$$

or, a weaker condition (for consistency):

$$E(x\epsilon) = 0 \tag{4}$$

Common Source of Estimation Bias

- Omitted Variable Bias
- Missing data and Sample Selection Bias
- Measurement errors
- Simultaneous causality bias

Omitted Variable Problem

$$y = \underline{x}'\underline{\beta} + c + u \tag{5}$$

- $\underline{x} = (1, x_2, \cdots, x_K)'$ – observable random variables.
- c – unobservable random variable.

- If $cov(x_j, c) = 0$ for all j, then we still get consistent estimator by simply ignoring c.
- However, if $cov(x_j, c) \neq 0$ for some j, putting c into the error term will cause some problem.

Omitted Variable Problem - Remedies

Remedies:

- Panel data methods (FE/RE)
- A randomized control experiment (DID)
- Instrumental Variables

Different types of data: Cross Sectional Data vs. Panel Data

Cross Sectional Data:

- Many different observations.
- Observed at a single point in time.
- N - sample size.

Different types of data: Cross Sectional Data vs. Panel Data

Cross Sectional Data:

- Many different observations.
- Observed at a single point in time.
- N - sample size.
- Survey data

  GDP, population, rate of unemployment, etc of all countries at the end of 2018.

Different types of data: Cross Sectional Data vs. Panel Data

Cross Sectional Data:

- Many different observations.
- Observed at a single point in time.
- N - sample size.

- Survey data

  GDP, population, rate of unemployment, etc of all countries at the end of 2018.

- Closing prices of a group of 20 different tech stocks on Dec 3rd, 2018.

Different types of data: Cross Sectional Data vs. Panel Data

Panel Data (Longitudinal Data):

- Repeated observations of the same cross section of units.
- T – # of time periods.
- T is usually small.

Different types of data: Cross Sectional Data vs. Panel Data

Panel Data (Longitudinal Data):

- Repeated observations of the same cross section of units.
- T – # of time periods.
- T is usually small.

  - GDP, population, rate of unemployment, etc of all countries for 2015, 2016, 2017, 2018

Different types of data: Cross Sectional Data vs. Panel Data

Panel Data (Longitudinal Data):

- Repeated observations of the same cross section of units.
- T – # of time periods.
- T is usually small.

- GDP, population, rate of unemployment, etc of all countries for 2015, 2016, 2017, 2018
- Closing prices of a group of 20 different tech stocks on Dec 3rd, 4th, 5th of 2018.

Unobserved Effects Models for Panel Data

For example, if we observe y and $\underline{x}$ at two time periods:

- $y_t, \underline{x}_t : t = 1, 2.$
- omitted variable c is time constant.

Then the population regression model:

- $y_t = \underline{x}_t' \underline{\beta} + c + u_t,\ t = 1, 2.$
- Note, c stays the same over time, but it is different for different individuals.

Unobserved Effects Models for Panel Data

$$y_t = \underline{x}_t' \underline{\beta} + c + u_t, t = 1, 2 \tag{6}$$

Case I:

- If $cov(\underline{x}_t, c) = \underline{0} \Rightarrow$ Ordinary Least Square (OLS).

Case II:

- If $cov(\underline{x}_t, c) \neq \underline{0} \Rightarrow$ OLS is inconsistent.
- We can obtain consistent estimator if we eliminate c.
- How can we do that? By taking the difference.
- $\Delta y = y_2 - y_1$; $\Delta \underline{x} = \underline{x}_2 - \underline{x}_1$; $\Delta u = u_2 - u_1$;

Unobserved Effects Models for Panel Data – Definition: Model

Unobserved effects model (UEM):

$$y_{it} = \underline{x}_{it}'\underline{\beta} + c_i + u_{it}, t = 1, \ldots, T. \tag{7}$$

- c_i is often referred to as unobserved component, latent variable or unobserved heterogeneity.
- It is also called individual effect/ individual heterogeneity.
- u_{it} is called the idiosyncratic errors/ idiosyncratic disturbances.

Unobserved Effects Models for Panel Data – Definition: Model

Based on the relationship between $\underline{x}_t$ and c, we have different names for the model

- $cov(\underline{x}_{it}, c_i) = \underline{0}$ – random effect model.
- $cov(\underline{x}_{it}, c_i) \neq \underline{0}$. – fixed effects framework.

  – individual fixed effect/ firm fixed effect model.

Example:

Program Evaluation: Model the effects of job training on subsequent wages:

$$\log(wage_{it}) = \theta_t + \underline{z}_{it}\nu + prog_{it}\delta_1 + c_i + u_{it} \tag{8}$$

- $\underline{z}_{it}$ includes the ith individual's characteristics.
- t = 1, no one has participated in the program: $prog_{i1} = 0$.
- t = 2, a subgroup has participated, so $prog_{i2} = 1$ for $i \in$ participants.

Unobserved Effects Models for Panel Data – Example

$$\log(wage_{it}) = \theta_t + \underline{z}_{it}\nu + prog_{it}\delta_1 + c_i + u_{it} \tag{9}$$

- c_i includes ability and $prog_{i2}$ might be affected by ability.
- For example, a person's choice to participate in a program might be the results of a person's ability – self-selection bias.

Example: Lagged Dependent Variable.

$$log(wage_{it}) = \beta_1 log(wage_{i,t-1}) + c_i + u_{it}, t = 1, \ldots, T. \tag{10}$$

- Interest lies in how does wage change after controlling for unobserved heterogeneity (individual productivity), c_i.
- Let $y_{it} = log(wage_{it})$.
- c_i must be correlated with $y_{i,t-1}$ thus x_{it}.

Unobserved Effects Models for Panel Data – Random Effect Model

Rewrite the model as

$$y_{it} = \underline{x}_{it}\underline{\beta} + v_{it}, \tag{11}$$

- $v_{it} \triangleq c_i + u_{it}$ – the composite errors.
- We can simply run the OLS estimation.
- To get consistent estimator, we should have Assumption RE.1:
  (1) $E(\underline{x}_{it} u_{it}) = \underline{0}$,
  (2) $E(\underline{x}_{it} c_i) = \underline{0}$.
- The key assumption here is (2).

Unobserved Effects Models for Panel Data – Random Effect Model

$$\underline{y}_i = X_i\underline{\beta} + \underline{v}_i, \underline{v}_i = c_i\underline{1}_T + \underline{u}_i \tag{12}$$

- $\Omega \triangleq E(\underline{v}_i\underline{v}_i')$ – T by T matrix.
- Assumption RE.2: $rankE(X_i'\Omega^{-1}X_i) = K$.

There is some structure for the error term in the RE model.

- How to make use of the known strucutre?

Let's assume $E(u_{it}^2) = \sigma_u^2$ and $E(u_{it}u_{is}) = 0$ for any $t \neq s$. Then:

- $E(v_{it}^2) = E(c_i^2) + 2E(c_iu_{it}) + E(u_{it}^2) = \sigma_c^2 + \sigma_u^2$.
- $E(v_{it}v_{is}) = E\{(c_i + u_{it})(c_i + u_{is})\} = \sigma_c^2$.

Unobserved Effects Models for Panel Data – Random Effect Model

So:

$$\Omega = E(\underline{v}_i \underline{v}_i') = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \dots & \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c^2 & \sigma_c^2 & \dots & \sigma_c^2 + \sigma_u^2 \end{pmatrix}$$

- The covariance matrix has a special structure.
- Also, we can write it as $\Omega = \sigma_u^2 I_T + \sigma_c^2 1_T 1_T'$.
- We say Ω has the random effects structure: It only depends on two parameters.

Estimation:

- $\hat{\Omega} \triangleq \hat{\sigma}_u^2 I_T + \hat{\sigma}_c^2 1_T 1_T'$.
- $\hat{\sigma}_v^2 = \frac{1}{NT-K} \sum_i \sum_t \check{v}_{it}^2$.
- $\hat{\sigma}_c^2 = \frac{1}{NT(T-1)/2-K} \sum_{i=1}^N \sum_{t<s} \check{v}_{it} \check{v}_{is}$

$$\underline{\hat{\beta}}_{RE} = (\sum_{i=1}^N X_i' \hat{\Omega}^{-1} X_i)^{-1} (\sum_{i=1}^N X_i' \hat{\Omega}^{-1} \underline{y}_i) \tag{13}$$

Unobserved Effects Models for Panel Data – Random Effect Model

To summarize, random effect estimator (RE estimator) is

$$\underline{\hat{\beta}}_{RE} = (\sum_{i=1}^{N} X_i' \hat{\Omega}^{-1} X_i)^{-1} (\sum_{i=1}^{N} X_i' \hat{\Omega}^{-1} \underline{y}_i) \tag{14}$$

- Under RE.1 and RE.2 $\underline{\hat{\beta}}_{RE}$ is consistent.
- $cov \underline{\hat{\beta}}_{RE} = E(X_i' \Omega^{-1} X_i)^{-1} / N$, can be estimated by $(\sum_{i=1}^{N} X_i' \Omega^{-1} X_i)^{-1}$.

Unobserved Effects Models for Panel Data – Random Effect Model

Test for $H_0 : \sigma_c^2 = 0$.

- First find an estimator of σ_c^2 and then find the limiting distribution under H_0.
- $\hat{\sigma}_c^2 = \frac{1}{NT(T-1)/2-K} \sum_{i=1}^{N} \sum_{t<s} \check{v}_{it} \check{v}_{is}$
- $\frac{1}{\sqrt{N}} \sum_i \sum_{t<s} v_{it} v_{is} \to \mathcal{N}(0, E(\sum_{t<s} v_{it} v_{is})^2)$
- Test Statistics (t - test):

$$\frac{\sum_{i=1}^{N} \sum_{t<s} \check{v}_{it} \check{v}_{is}}{\{\sum_{i=1}^{N} (\sum_{t<s} \check{v}_{it} \check{v}_{is})^2\}^{1/2}} \tag{15}$$

Unobserved Effects Models for Panel Data – Random Effect Models

Example (RE Estimation of the effects of Job traning grants)

$$\log(scrap)_{it} = \beta_0 + \beta_1 I(1988) + \beta_2 I(1989) + \beta_3 grant + \beta_4 grant_{-1} + c_i + u_{it}$$

- Estimation using RE model.
- Test for whether $\sigma_c^2 = 0$.

Rcode available at RE_jtrain.R

Unobserved Effects Models for Panel Data – Fixed Effect Model

Idea: instead of putting c_i into the error term, we want to treat it as a parameter.

$$\underline{y}_i = X_i \underline{\beta} + c_i \underline{1}_T + \underline{u}_i \tag{16}$$

- Assumption FE.1: $E(u_{it}|\underline{x}_i, c_i) = 0, t = 1, \ldots, T$.
- $\underline{1}_T$ is a vector of 1, of length T.
- Because of this, x_{it} cannot contain any time invariant variable: so it is often called time varying explanatory variable.
- For example, if we have a panel of adults and one element of x_{it} is education. While it can be constant for some part of the sample, we must have education changing for some people in the sample.

Unobserved Effects Models for Panel Data – Fixed Effect Model

How do we do the estimation?

- First, calculate the average over time for each unit.
- $\bar{y}_i = \frac{1}{T}\sum_t y_{it}$, $\underline{\bar{x}}_i = \frac{1}{T}\sum_t \underline{x}_{it}$, $\bar{u}_i = \frac{1}{T}\sum_t u_{it}$.
- $\bar{y}_i = \underline{\bar{x}}_i'\underline{\beta} + c_i + \bar{u}_i$ – between equation.
- $\underbrace{y_{it} - \bar{y}_i}_{\ddot{y}_{it}} = \underbrace{(\underline{x}_{it}' - \underline{\bar{x}}_i')}_{\ddot{\underline{x}}_{it}}\underline{\beta} + \underbrace{u_{it} - \bar{u}_i}_{\ddot{u}_{it}}$ – within equation.
- This is called the fixed effect transformation or within transformation.

Unobserved Effects Models for Panel Data – Fixed Effect Model

$$\ddot{y}_{it} = \underline{\ddot{x}}'_{it}\underline{\beta} + \ddot{u}_{it} \tag{17}$$

- Estimation can be done using OLS (also called pooled OLS).
- It is easy to verify $E(\underline{\ddot{x}}_{it}u_{it}) = \underline{0}$.
- We call the pooled OLS estimators $\underline{\hat{\beta}}_{FE}$ the FE estimator.
- Assumption FE.2: $rank(\sum_t E(\underline{\ddot{x}}_{it}\underline{\ddot{x}}'_{it})) = rank\{E(\ddot{X}_i'\ddot{X}_i)\} = K$.

Unobserved Effects Models for Panel Data – Fixed Effect Model

$$\underline{\hat{\beta}}_{FE} = (\sum_i \ddot{X}_i' \ddot{X}_i)^{-1} (\sum_i \ddot{X}_i' \underline{\ddot{y}}_i) = (\sum_i \sum_t \underline{\ddot{x}}_{it} \underline{\ddot{x}}_{it}')^{-1} (\sum_i \sum_t \underline{\ddot{x}}_{it} \ddot{y}_{it}) \quad (18)$$

- The above is called the within estimator.
- FE.1 - FE.2 will ensure consistent estimator.

Unobserved Effects Models for Panel Data – Fixed Effect Model

- Assumption FE.3 $E(\underline{u}_i\underline{u}_i'|\underline{x}_i, c_i) = \sigma_u^2 I_T$.
  $\Rightarrow E(u_{it}u_{is}|\underline{x}_i, c_i) = 0$ and $E(u_{it}^2|\underline{x}_i, c_i) = \sigma_u^2$.
- Under FE.1 - FE.3, we can get:

$$\sqrt{N}(\underline{\hat{\beta}}_{FE} - \underline{\beta}) \xrightarrow{d} N(\underline{0}, \sigma_u^2[E(\ddot{X}_i'\ddot{X}_i')]^{-1})$$

- $\hat{\sigma}_u^2 = \frac{SSR}{N(T-1)-K} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\ddot{u}}_{it}^2}{N(T-1)-K}$

Unobserved Effects Models for Panel Data – Fixed Effect Model

Example: (FE Estimation of the effects of Job traning grants)

$$\log(scrap)_{it} = \beta_0 + \beta_1 I(1988) + \beta_2 I(1989) + \beta_3 union_i \\ + \beta_4 grant + \beta_5 grant_{-1} + c_i + u_{it}$$

- Estimation using FE model.
- Cannot include intercept, *union* as explanatory variables

Rcode available at FE_jtrain.R

Unobserved Effects Models for Panel Data – Random Effect model vs Fixed Effect model:

Random Effect model vs Fixed Effect model:

- If $E(c_i \underline{x}_{it}) = \underline{0}$, then it's better to use RE. Since RE estimators can have much smaller variance.
- If some key variables in $\underline{x}_{it}$ do not change much over time, then FE will get imprecise estimates.
- Also, you should think of what do you want to predict: do you want to predict another firm or one of the firm in your sample.

Test for $E(c_i \underline{x}_{it}) = \underline{0}$

- Hausman Test. Idea: when $E(c_i \underline{x}_{it}) = \underline{0}$ is true, both RE and FE will provide consistent estimator.

Unobserved Effects Models for Panel Data – Random Effect model vs Fixed Effect model:

From panel data to data with structure.

The Difference-in-Differences Estimator

Similar Motivation:

- No luxury of random samples.
- Selection bias – the selection into control/treatment group is by choice

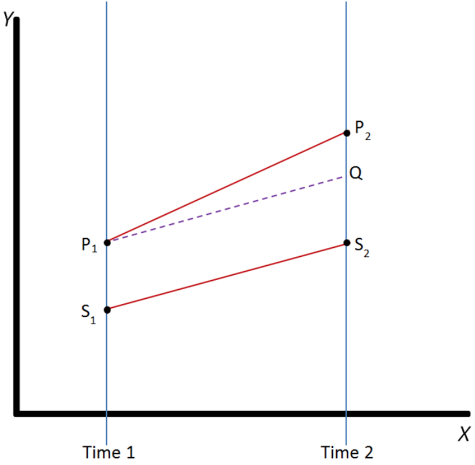
Minimum wage law example:

- Y: fte (full-time equivalent employment)
- X: minimum wage increase, NJ vs non NJ, before/after the law.
- obs: a collection of resturant before/after the law.

References:

- https://bookdown.org/ccolonescu/RPoE4/indvars.html#the-difference-in-differences-estimator

The Difference-in-Differences Estimator



Source: https://en.wikipedia.org/wiki/Difference_in_differences

The Difference-in-Differences Estimator

Four averages of the response:

- $\bar{y}_{T,A}$ – treatment, after
- $\bar{y}_{C,A}$ – control, after
- $\bar{y}_{T,B}$ – treatment, before
- $\bar{y}_{C,B}$ – control, before

The difference-in-differences estimator $\hat{delta}$ is defined as

$$\hat{delta} = (\bar{y}_{T,A} - \bar{y}_{T,B}) - (\bar{y}_{C,A} - \bar{y}_{C,B}) \tag{19}$$

An equivalent regression model:

$$y_{it} = \beta_1 + \beta_2 T + \beta_3 A + \delta T \times A + e_{it} \tag{20}$$

Reference

https://www.econometrics-with-r.org/9-asbomr.html

- Ch 9: Source of estimation bias
- Ch 10: FE/RE
- Ch 13: DID

https://bookdown.org/ccolonescu/RPoE4/

- Ch 7.7 Diff-in-Diff
- Ch 15: Fixed/Random