Recitation 6: Assessing the models and Panel regression

Seung-hun Lee

Columbia University
Undergraduate Introduction to Econometrics Recitation

October 27th, 2022

Assessing the regression model

Critiquing the regressions

- **External validity** is concerned with the applicability of the regression model to other contexts.
  - For instance, you may wonder whether regression model from schools in California could explain what happens in the classrooms in South Korea
  - Any critiques to the model questioning the replicability of the result to other dataset is assessing the external validity.
- **Internal validity** assesses the model from the point of view of the population being studied.
  - This approaches questions the validity of the statistical inference within the given dataset. There are five threats to internal validity.

Is the model itself set properly?

- **Omitted variable bias:** Did the researcher left out factors that could affect Y and are correlated with X?
- **Wrong functional form:** Did the researcher incorporate X in a proper functional form? (quadratic, logs?)
- **Errors in variables bias:** Measurement error. If X does not have a clear cut measure, part of X correlated to Y and the observed version of X is left out, resulting in attenuation bias
- **Sample selection bias:** If certain groups are more likely to not respond, then the data fails to represent population and the estimates based on this data is biased.
- **Simultaneous causality bias:** There are cases where Y causes X too, also leading to the bias (consider supply and demand)

Dealing with omitted variable bias

- We know OVB is a problem if
  - that variable is a determinant of the dependent variable
  - that variable is correlated with a regressor already in the model
- Easy: Including the omitted variable itself or putting in relevant control variables
- Challenging ones
  - **Instrumental variables :** The idea is to find another variable (Z) that is relevant to the included regressor that may be endogenous and does not affect the outcome on its own (only affect outcome through its relationship with the included regressors).
  - **Panel regression :** Each entity is observed across multiple time periods. Control for unobservable time-invariant characteristics of the individual (Individual fixed effects) and characteristics that are common across all individuals in that period (Time fixed effects).

Wrong functional forms

- This concerns a case where the regression model is in an incorrect form (enforcing linearity when nonlinear regressor can do better)
- Can bias the estimates: Effects of X may not be estimated correctly
- Include a squared, log, or interacted variables in the equation
- A special case of this when a dependent variable is a binary variable. Instead of a linear regression model, we have to use a probit or a logit mode that allows for the parameters to be nonlinear with the regressors

Measurement error

- Our independent variables may not always be accurately measured
- Classical case: We observe $\tilde{X}_i$ instead of X_i, where $\tilde{X}_i = X_i + w_i$
- w_i term satisfies $E[w_i] = 0, cov(X_i, w_i) = 0, cov(u_i, w_i) = 0$
- OLS estimate is biased towards zero: Probability limit of OLS estimate is

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \right)$$

Attenuation bias

- Best guess error: Guess of X_i is precise conditional on observables ($\tilde{X}_i = E[X_i|W_i]$, where $E[u_i|W_i] = 0$)
  - No bias: We get $cov(\tilde{X}_i, u_i - \beta_1 w_i) = 0$ (Above, we have this equal to $-\sigma_w \beta_1$)
  - Rare: Guess has to be precise (no error at all)

Proof on classical Measurement Error

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad \tilde{X}_i = X_i + w_i$$

$\text{Cov}(X_i, u_i) = 0$

$E(w_i) = 0$

$\text{Cov}(X_i, w_i) = 0$

$\text{Cov}(u_i, w_i) = 0$

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + (u_i - w_i\beta_1)$$

v_i

$$\hat{\beta}_1 = \frac{\sum(\tilde{X}_i - \bar{\tilde{X}})(Y_i - \bar{Y})}{\sum(\tilde{X}_i - \bar{\tilde{X}})^2} = \beta_1 + \frac{\sum(\tilde{X}_i - \bar{\tilde{X}})(v_i - \bar{v})}{\sum(\tilde{X}_i - \bar{\tilde{X}})^2}$$

$\tilde{X}_i = X_i + w_i$, $\bar{\tilde{X}} = \bar{X} + \bar{w}$

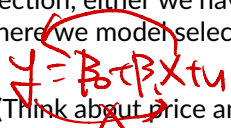
$$= \beta_1 + \frac{\sum[(X_i - \bar{X}) + (w_i - \bar{w})][(w_i - \bar{w}) - (w_i - \bar{w})\beta_1]}{\sum[(X_i - \bar{X}) + (w_i - \bar{w})]^2}$$

$$= \beta_1 - \frac{\sum(w_i - \bar{w})^2\beta_1}{\sum(X_i - \bar{X})^2 + \sum(w_i - \bar{w})^2} = \beta_1\left(1 - \frac{\sum(w_i - \bar{w})^2}{\sum(X_i - \bar{X})^2 + \sum(w_i - \bar{w})^2}\right)$$

$$= \beta_1 \frac{\sum(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2 + \sum(w_i - \bar{w})} = \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$$

Selection bias and simultaneous causality

- Selection bias
  - Does our sample match the population of interest?
  - Missing data: Not too worrisome if missing at random or based on some observable X (include X as controls)
  - Missing based on Y: Sample is selected based on outcome, signaling bias in estimates.
  - In order to address selection, either we have to adjust the population of interest or have a separate regression where we model selection into treatment
- Reverse causality
  - IRL, Y might cause X (Think about price and quantity)
  - We need to model for exogenous changes that affects one part of the causal chain but not the other using instrumental variables or explicitly model for both causations using structural approach



Panel regression model

Motivation and advantages for panel estimation

- **Panel data**: We observe multiple individuals for multiple periods of time.

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

$i = 1, 2, ..., N \rightarrow$ individuals, $t = 1, 2, ..., T \rightarrow$ time periods.

- Balanced: There are T datasets for each of the N individuals.
- Unbalalced: There are $t \leq T$ datapoints for some of the N individuals.

- Panel data allows us to use more datasets.
- Panel data allows us to control for **unobserved heterogeneity** that are
  1. different accross N entities but always remain same for T periods in a given entity (**cross section fixed effect**)
  2. different accross T times periods but remains the same for all N entities in a particular time period (**time fixed effects**)
  3. both of 1) and 2). (**two-way fixed effects**)

Interacted fixed effects

What OVB problems could we be dealing with?

- 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

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

fatality rate → $\hat{Y}_{i1}$; X_{i1} → beer tax; (0.20) → not significant at all; (0.20) → significant positive

- In such case, one might suspect that there is an omitted variable bias that affects these coefficients.
  - Omitted variable specific to the states (Strictness of the relevant law)
  - Time-trends? (Specific to each of years 1 and 2)

How can panel regression do better?

- Let Z_i denote the strictness of state laws on DUI that are unchanging.
- Now write

$$Y_{i1} = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_i + u_{i1}$$
$$Y_{i2} = \beta_0 + \beta_1 X_{i2} + \beta_2 Z_i + u_{i2}$$

- 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 - because it is now gone.
- If we estimate this particular β_1, we can obtain much more accurate estimates of the effect of beer tax on fatality rate.

Specific methodologies for cross-sectional FE

- There are two ways of estimating the data when $T \geq 3$
- Least square dummy variables (LSDV): Include $N - 1$ individual dummies
- Within estimation: Subtract "demeaned" equation from the original
- Use:

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

where Z_i is the cross section fixed effect.

- Define $\alpha_i = \beta_0 + \beta_1 Z_i$. Then the above equation can be written as

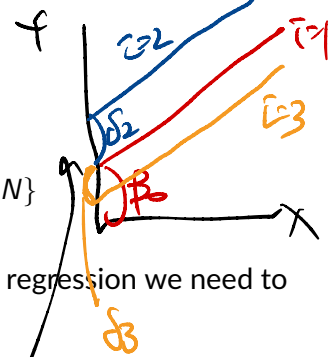
$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \tag{2}$$

- α_i term can be thought of as an effect of being an entity i, which is **correlated with** X_{it}

LSDV method

- Define a new variable D_{ki} as follows

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



- Since we are going to include β_0, a common intercept, in our regression we need to remove one of the N (dummy variable trap)
- Then we can write

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

- This equation gives different intercepts for each i (can you see why?), while keeping the slope on X_{it} constant at β_1
- Control for unobserved cross section fixed effect by allowing the intercept to differ by each i

Within estimation methods

- Define $\bar{X}_i, \bar{Y}_i$ as sample mean of X_{it}, Y_{it} for given i over all possible t's.

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

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

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

- Subtract Y_{it} by $\bar{Y}_i$ to get

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

- This process gets rid of α_i. Then, apply OLS estimation on this equation to get the within estimator

Having both FEs with two-way fixed effects

- We have a DGP

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

- LSDV: With an overall constant β_0, we can put $N - 1$ individual and $T - 1$ time dummies
- WE: Demeaning should be done in the following method

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

This (and only this) would allow us to get rid of both the α_i individual fixed effect and the λ_t time effects

Least square assumptions for panels

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.

P2 : $(X_{i1}, .., X_{iT}, u_{i1}, ...u_{iT})$ is IID across $i = 1, .., n$. **This does not rule out the correlation between u_{it}, u_{ij} within entity i for different j and t**, allowing serial correlation within the same entity

P3 : (X_{it}, u_{it}) have nonzero finite fourth moments (outliers are very unlikely) so that the panel estimators have a distribution

P4 : There is no perfect multicollinearity

P5 : X_{it} varies across time (for $\hat{\beta}_1$ to be defined)

$\rightarrow$ Because of P2, we need to use **clustered standard error** at a cross-sectional level.