Introduction to Econometrics: Recitation 6

Seung-hun Lee*

October 27th, 2022

1 Assessing the model: Internal/exteral validity

Not all regression models are perfect. The result may only apply to limited context. There may be factors within the regression model that prevents the model from delivering accurate verdict.

External validity is concerned with the applicability of the regression model to other contexts. This requires the deep understanding of the case being studied by the model and how they are similar/different from other cases. 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. Specifically, this approaches questions the validity of the statistical inference within the given dataset. There are five threats to internal validity.

- **Omitted variable bias:** Whenever the regression model contains control variables, we should be concerned about whether the researcher left out factors that could affect Y and are correlated with X. If so, the error term and X is correlated and the estimates are biased.
- **Wrong functional form:** This occurs when the researcher does not include X properly. Perhaps it is a mistake to contain X alone. One can suggest including X in a quadratic form or with some interaction term.

*Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

- **Errors in variables bias:** Also known as measurement error. Perhaps our variable X does not have a clear cut measure. If so, some part of X which is related to Y and the observed version of X is not included, resulting in an attenuation bias.
- **Sample selection bias:** Consider a survey. An ideal survey would represent various groups in the population. Now, let's say people don't respond anymore. If certain groups are more likely to not respond, then the survey fails to represent population and the estimates based on this data is biased.
- **Simultaneous causality bias:** We assumed that X causes Y. However, there are cases where Y causes X too. We assumed that class size, our X causes some result in test scores Y. However, it is possible that after observing smaller size class performs well, schools start reducing class size. In this case, our β coefficient will be biased.

1.1 Dealing with omitted variable bias

We have learned that omitted variable can cause problem if 1) that variable is a determinant of the dependent variable and 2) that variable is correlated with a regressor already in the model. We have also talked about an easy solution: Including the omitted variable itself or putting in relevant control variables that are relevant to the omitted variables. There are more technical solutions that we will cover as a separate chapter

- **Instrumental variables :** We have to resort to this when there is no way to capture the omitted variable. The idea is to find another variable (commonly denoted as Z) that is relevant to the included regressor that may be endogenous and does not affect the outcome on its own (or only affect outcome through its relationship with the included regressors).
- **Panel regression :** Each entity is observed across multiple time periods. If you include dummies for each individuals, you can control for unobservable characteristics of the individual that do not change over time (Individual fixed effects). You can also include dummies for time periods that represent characteristics that are common across all individuals in that period (Time fixed effects).

1.2 Wrong functional form

This concerns a case where the regression model is in an incorrect form, like only having linear terms where a squared or a log of a variable might do a better job. This error, also

known as misspecification error, may lead to a bias in estimate. To address this, you can 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.

1.3 Measurement errors

There might be a case where the variable X is measured with an error - due to inaccurate memory of the respondent, questions that are poorly framed or ask too much details, and simple data entry. The problem with the classical measurement error in the independent variable is that it will drive the estimates towards zero (an attenuation bias)

Here is how, suppose that the true model is

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

where X_i is the true variable of interest. Instead of X_i, we have an $\tilde{X}_i$ which is defined as

$$\tilde{X}_i = X_i + w_i$$

where w_i term satisfies $E[w_i] = 0, cov(X_i, w_i) = 0, cov(u_i, w_i) = 0$. So replace X_i with $\tilde{X}_i - w_i$ and do an OLS, which gets us

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + u_i - w_i\beta_1 \rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n (\tilde{X}_i - \bar{\tilde{X}}_i)(Y_i - \bar{Y}_i)}{\sum_{i=1}^n (\tilde{X}_i - \bar{\tilde{X}}_i)^2}$$

This can be re-written as

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n [(X_i - \bar{X}_i) + (w_i - \bar{w}_i)][(u_i - \bar{u}_i) + (w_i - \bar{w}_i)\beta_1]}{\sum_{i=1}^n [(X_i - \bar{X}_i) + (w_i - \bar{w}_i)]^2} \\
&= \beta_1 - \beta_1 \frac{\sum_{i=1}^n (w_i - \bar{w}_i)^2}{\sum_{i=1}^n [(X_i - \bar{X}_i) + (w_i - \bar{w}_i)]^2} \\
&= \beta_1 \left(1 - \frac{\sum_{i=1}^n (w_i - \bar{w}_i)^2}{\sum_{i=1}^n [(X_i - \bar{X}_i) + (w_i - \bar{w}_i)]^2}\right) = \beta_1 \left(1 - \frac{\sigma_w^2}{\sigma_X^2 + \sigma_w^2}\right) = \beta_1 \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\right)
\end{aligned}$$

Thus, the OLS estimate is biased towards zero.

Here is another measurement error: a best-guess measurement error. This is when a respondent does not accurately remember the true X_i but makes a guess based on some W_i,

such that $\tilde{X}_i = E[X_i|W_i]$, where $E[u_i|W_i] = 0$. Put it differently, we have $w_i = X_i - \tilde{X}_i = X_i - E[X_i|W_i]$. In this case, we have the following for $cov(\tilde{X}_i, u_i - \beta_1 w_i)$

$$\begin{aligned} cov(\tilde{X}_i, u_i - \beta_1 w_i) &= cov(E[X_i|W_i], u_i - \beta_1 w_i) \\ &= cov(E[X_i|W_i], u_i) - cov(E[X_i|W_i], \beta_1(X_i - E[X_i|W_i])) \\ &= E[E[X_i|W_i]u_i] - E[E[X_i|W_i]]E[u_i] \\ &- E[E[X_i|W_i](X_i - E[X_i|W_i])\beta_1] + E[E[X_i|W_i]]E[X_i - E[X_i|W_i]]\beta_1 \\ &= 0 \end{aligned}$$

by applying the law of iterated expectations on the last two lines, each term is equal to 0. So this type of error leads to no bias. However, this is an extreme case in that you need to make the best guess where the guess conditional on observables are precise.

1.4 Selection bias

This occurs in case where we have a justifiable reason to guess that our sample is significantly different from the population of interest. One possible case of this is when we have a missing data problem. When the data is missing at random or based on the value of independent variables, it does not introduce a bias (standard errors may rise though) and all we have to do is to include control variables for this. When the data is missing due to the value of our dependent variable, we have a problem. This means that the analysis sample is selected based on outcome, suggesting that not taking this into account leads to estimation bias. 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 (For technical details, refer to Lee (2009) in *Review of Economic Studies*).

1.5 Simultaneous causation bias

We have assumed that X causes Y but not the other way around. However, there are many cases in reality where we cannot rule out the causation going the other direction. Think about the regression where we have quantity on the left and price on the right hand side. Here, causation can run two ways - changes in price affect quantity demanded (supplied), but it might be that changes in quantity affects price too. To address this, 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.

2 Panel Regression

2.1 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

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

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).

2.2 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

2.2.1 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

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 D_{2i} + ... + \delta_N D_{Ni} + u_{it} \qquad (3, \text{LSDV})$$

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

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

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.

2.2.2 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

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

2.2.3 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.

2.2.4 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

P5 : X_{it} should vary across time. This variation is part of the equation for determining β_1. If X_{it} does not vary across time, β_1 will be unestimable.

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.