ECON 3040 - Panel Data

Ryan T. Godwin

University of Manitoba

Missing variables (unobserved heterogeneity)

- We return to the same problem that we studied in Instrumental Variables: the dreaded **missing variable** problem. Recall that if a variables is missing, and it's correlated to x, then the LS estimator for the effect of x on y is biased and inconsistent.
- In panel data, we sometimes call the missing variables problem **unobserved heterogeneity** or **unobserved effects**.
- If we can observe the same units of observations (e.g. individuals or countries) repeatedly over time, then we can solve this unobserved heterogeneity problem, using the **fixed effects** or **within** estimator.

What is panel data?

- "Panel data" is a data set where we observe the same observational units (individuals, countries) repeatedly over time. Each "panel" represents a moment in time, a snapshot of all the individuals in the sample. Then, this panel of information is repeated over time, with some or all of the values of the variables changing in each panel.
- A panel data set can be arranged by "stacking" each person's information over time horizontally, and then going on to the next person. Alternatively, we could add another set of n observations at the bottom of the data set for each time period, but where the individuals in each stack are the same people.
- In a panel data set we have $n \times T$ observations. n are the number of individuals, and T are the number of time periods over which these individuals are observed.
- We will need to introduce a new subscript in our notation: $t = 1, \ldots, T$. The dependent variable is now written as y_{it}.

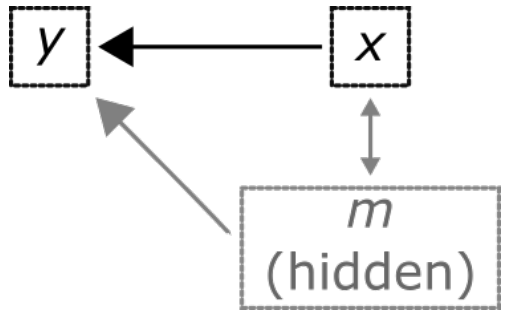
The model, over time

The population model, with the missing variable, now looks something like:

$$y_{it} = \beta_0 + \beta_1 x_{it} + m_i + \epsilon_{it},$$

- m_i is the effect the missing variable has on y_{it}
- m_i has an i subscript, but no t subscript. This means that it varies by individual, but not over time.
- m_i is called *unobserved heterogeneity* because it accounts for differences in each person. But these characteristics do not change over time.

Figure: A missing m variable that is correlated with x and that determines y will make estimation of the effect of x on y difficult (or impossible).



Let's redraw this and add in the *subscripts*.

LS estimation with panel data

There are several things we can do with panel data. The most basic is called **pooled OLS**. We simply ignore the panel structure and "pool" all the data together. This gives us $n \times T$ observations. But we can actually use panel data to solve the missing variable problem (and other problems of endogeneity), by using the *fixed effects* estimator.

m_i is time invariant

If the missing variable m_i does not vary over time, but if the y_{it} and x_{it} variables *do* vary over time, then we can use a trick to solve the missing variable (unobserved heterogeneity) problem. y_{it} could be an individuals wages, for example, and x_{it} could be some economic policy or some choice that people are making, that changes over time. Some examples of m_i (unobserved heterogeneity) that we will be able to deal with, are:

- gender
- race
- ability

These are things that are specific to each individual, but that (usually) do not vary over time. If these things are unobserved, we can solve the problem using the **fixed effects** estimator!

Wages and marriage

It is a stylized fact that married people earn more than single people (that means that we observe a positive correlation in wage data, married workers have higher than average wages). Does marriage **cause** wages? What could be some possible explanations for this association?

$$wage_{it} = \beta_0 + \beta_1 marriage_{it} + \beta_2 x_{it} + \beta_3 x_i + ability_i + \epsilon_{it},$$

- $ability_i$ is a missing variable that determines wage.
- We can't estimate β_1 and call it the causal effect of marriage on wage. Why?

Suppose we observe the workers' wages and marital status over time. We also have some time-varying controls (x_{it} - things like the amount of time worked that year) and some time invariant controls (x_i - like gender or race). It turns out that the x_i variable can't be used; it will disappear when we make the missing $ability_i$ variable cancel out.

The trick

$$y_{it} = \beta_0 + \beta_1 x_{it} + m_i + \epsilon_{it},$$

- What happens if something doesn't change over time, and you take it's average over time?
- How can we get rid of (subtract) the missing variable (the individual heterogeneity)?

The trick involves taking the averages of the variables *over time*. That is, we're going to take each person's wage, and average it over the time in which they are observed. Let's define these time-averages:

$$\bar{y}_i \equiv \frac{1}{T} \sum_{t=1}^{T} y_{it}$$

$$\bar{x}_i \equiv \frac{1}{T} \sum_{t=1}^{T} x_{it}$$

The trick

- If we could see the missing variable (we can't), what would be it's average over time? Does it change?
- What about β_0?
- What about the error term?

$$\frac{1}{T}\sum_{t=1}^{T} m_i = m_i$$

$$\frac{1}{T}\sum_{t=1}^{T} \beta_0 = \beta_0$$

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} = \bar{\epsilon_i}$$

So anything that is constant over time, is the same as it's average over time.

The trick

$$y_{it} = \beta_0 + \beta_1 x_{it} + m_i + \epsilon_{it} \tag{1}$$

Let's instead write the population model in terms of the over-time-averages:

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

Can you see a way to delete the missing variable? Subtract (2) from (1)!

$$(y_{it} - \bar{y}_i) = (\beta_0 - \beta_0) + (\beta_1 x_{it} - \beta_1 \bar{x}_i) + (m_i - m_i) + (\epsilon_{it} - \bar{\epsilon}_i)$$

$$(y_{it} - \bar{y}_i) = (\beta_1 x_{it} - \beta_1 \bar{x}_i) + (\epsilon_{it} - \bar{\epsilon}_i)$$

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{\epsilon}_{it}$$

Time-demeaning deletes the unobserved heterogeneity

- When we subtract the time-averages (time demeaned variables) from both sides of the equation, we are deleting the unobserved heterogeneity (missing variable).
- The double-dots denote a time demeaned variable, for example $\ddot{y}_{it} = y_{it} - \bar{y}_i$
- Note that any individual characteristics that are time invariant, such as race, gender, etc., will also be deleted from the regression. That is why x_i is not needed as a control, only x_{it}.
- We can estimate this model in R by manually de-meaning the variables, and estimating the model by LS. This is called the **within** estimator, which is also equivalent to the fixed effects estimator.

Fixed effects estimation

- The "within" estimator is more commonly called the **fixed effects** estimator.
- Remember how dummy variables measure the difference in group means?
- If we give each person their own dummy variable, called a **fixed effect**, and estimate the model with these individual dummies, we get the same thing as the **within** estimator!

That is, we estimate the equation:

$$y_{it} = \beta_1 x_{it} + \alpha_i D_i + \epsilon_{it},$$

Where α_1 is person 1's coefficient, and D_1 is a dummy variable equal to 1 if the observation is for person 1, and 0 otherwise. Including these individual fixed effects effectively demeans the data and eliminates the unobserved heterogeneity (missing variable) problem.