**Econometrics**

Chapter 10: Panel Data

If we have data on observations over time, it gives us one way to deal with OVB.

Specifically if the unobserved variable that is causing the problem is constant over time, then we can control for it even if we do not observe it.

Panel Data (longitudinal): data on $n$ entities over $T$ time periods.

Example: may have data on 48 states (n = 48) over a 7 year period (T = 7) for a total of 7x48 = 336 observations.

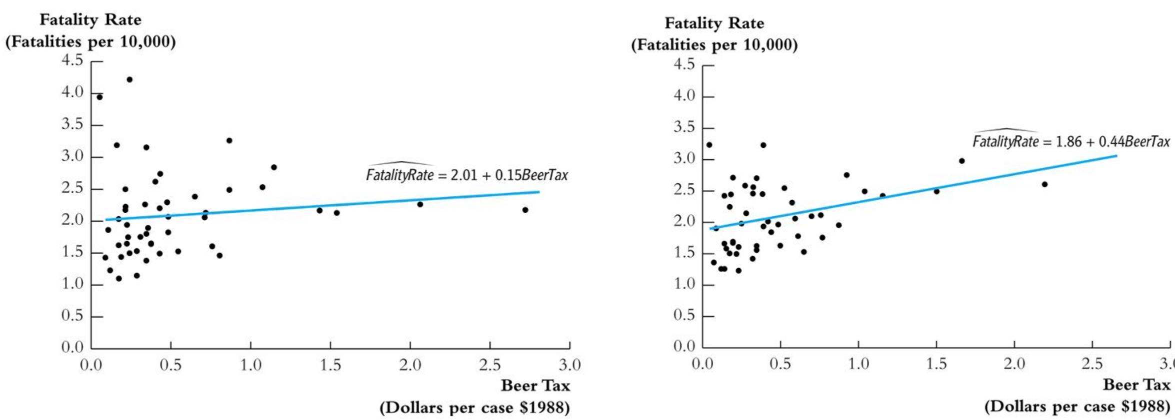
We generally denote the observation first, then the time period, in the subscripts: $Y_{it}$.

So $Y_{4,5}$ is the 4th entity in the 5th time period.

If all observations are available for each time period we say we have a **balanced panel**.

Lets look at traffic deaths and alcohol taxes in the US.

<u>Variables:</u>
Traffic deaths: # deaths in state in a year per 10,000 residents
Tax on bear per case
Others Vars.: driving age, drinking laws etc.




(a) 1982 Data

(b) 1988 Data

Figure 1: Higher Taxes More Deaths??

Why would there be more deaths in states with higher taxes?

A: Traffic Density -
Western states have lower density and also tend to have lower taxes then Eastern states. This is just a case of omitted variable bias then.

B: Attitudes about drinking -
The culture of a state probably affects traffic deaths and the tax rates on alcohol. Again this is an omitted variable problem.

*Panel Data with Two Time Periods*

Since we have two time periods we can look at changes in the $Y$ variable.

And since anything that is constant across time - like maybe traffic density and attitudes not changing quickly - will be the same in the two periods, they cannot be an 'omitted variable problem' - cannot affect the *change* in Y.

Consider some variable $Z$ that affects fatalities, but does not change over time (or very slowly).

$\hat{FatalityRate}_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}$

Note the subscript on $Z$ does not have a 't' because it is not changing over time.

$\hat{FatalityRate}_{i,1982} = \beta_0 + \beta_1 BeerTax_{i,1982} + \beta_2 Z_i + u_{i,1982}$
$\hat{FatalityRate}_{i,1988} = \beta_0 + \beta_1 BeerTax_{i,1988} + \beta_2 Z_i + u_{i,1988}$

Now subtract one from the other:

$FatalityRate_{i,1988} - FatalityRate_{i,1982} = \beta_1(BeerTax_{i,1988} - BeerTax_{i,1982}) + (u_{i,1988} - u_{i,1982})$

Note that our unobserved $Z$ that was causing problems is now gone - no more OVB.

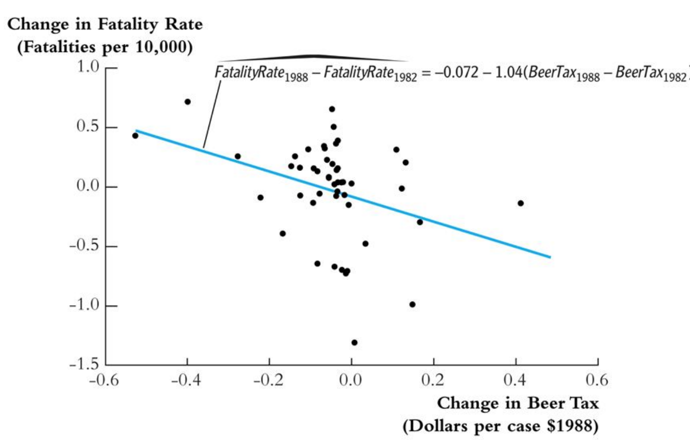



Figure 2: Two Time Period Panel

Note there is an intercept included which allows for the change in fatalities to be non-zero even if there is no change in beer tax.

Now the SE on $\beta_1$ here is 0.36, so we can reject a zero effect of changing beer tax on fatalities.

And this is what we would expect - increasing beer tax reduces fatalities.

Now, what if we have more than only two time periods: $T > 2$?

*Fixed Effects Regression*

Recall that in general what we have is:

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

But since $Z$ does not change over time, we can just interpret this as a single model with $n$ intercepts - one for each entity.

We let $\alpha_i = \beta_0 + \beta_2 Z_i$ and get:

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

This is called a fixed effects regression model - each entity has its own intercept or 'fixed effect' ($\alpha$

So the slope is the same for all entities, but each has a different intercept.

We can also think about this as just a function of $Y$ on $X$ along with an intercept and $n-1$ dummy variables, where each dummy represents a state/entity. (Make sure you see this - write it out and check)

And of course we can easily extend this model to include multiple $X$s.

Now, in theory we could just run OLS on this big model and each intercept captures any possible variables causing OVB.

But there are $k+n$ coefficients to estimate. In some cases this can be hard to solve. So in practice something else is done.

*Demeaned OLS*

First recall what we are working with:

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

Now take the average of both sides (over T's):

$\overline{Y}_i = \beta_1 \overline{X}_i + \alpha_i + \overline{u}_i$ where $\overline{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it}$ and same for $X$ and $u$

This implies: $Y_{it} - \overline{Y}_i = \beta_1 (X_{it} - \overline{X}_i) + (u_{it} - \overline{u}_i)$

So if we let $\tilde{Y}_{it} = Y_{it} - \overline{Y}_i$. And similarly define $\tilde{X}_i$ and $\tilde{u}_i$. We have:

$\tilde{Y}_i = \beta_1 \tilde{X}_i + \tilde{u}_i$

And so we can just run OLS on our demeaned data and recover our $\beta_1$

*Assumptions of Fixed Effect Regression and SEs*

1. $E(u_{it}|X_{i1}, X_{i2}, ...., X_{iT}, \alpha_i) = 0$
So each error term is mean independent of *all* of the Xs for that entity. This implies there is no OVB.

2. $(X_{i1}, X_{i2}, ..., X_{iT}, u_{i1}, u_{i2}, ...., u_{iT}), i = 1, 2, ..., n$ are i.i.d draws from their joint distribution. Note this says that the set of regressors and errors for one entity is independent, but identically distributed, to the other entities.

3& 4 are the same as always - finite fourth moment and no perfect multicollinearity.

Note how assumption 2 differs than for OLS for cross-section data. For cross-section each *observation* is independent, but here multiple observations for the same entity are allowed to be dependent, just not between entities.

When a variable is correlated with itself over time we say it is autocorrelated or serially correlated.

If we sample our entities by simple random sampling, then $(u_{i1}, u_{i2}, ...u_{iT})$ is independent of $(u_{j1}, u_{j2}, ...u_{jT})$, but in general $u_{it}$ will be serially correlated.

Under our 4 assumptions our OLS $\beta$s are unbiased, consistent, and asymptotically normal.

But, we need a new form for appropriate SEs.

We need ones that are heteroskedastic and autocorrelation-consistent, thus they are called HAC standard errors.

The ones we will use are clustered SEs. That is they allow the error terms to have any arbitrary autocorrelation within a cluster (the entities) but are assumed uncorrelated across entities.