Panel Data —Chapter 14 of Wooldridge's textbook

(Jing Li, Miami University)

Big Picture

In this lecture you will learn

1. Panel Data
2. One Way Fixed Effect Model
3. Two Way Fixed Effect Model
4. Cluster Robust Standard Error

Main Ideas

1. Panel data is obtained by observing the same person, firm, county, state, etc over several periods
2. Panel data can be used to control for unobserved factors or confounders that vary across units but are constant over time
3. Panel data can also control for certain time-varying confounders
4. The key idea is using dummy variables as proxy for those confounders
5. Observations for the same unit (entity, panel, cluster) in general are dependent. Thus cluster-robust statistics that account for correlation within cluster should be used

Organizing Panel Data

1. It is important to have an ID variable that distinguishes one unit from others, such as patient ID, firm ID and state name
2. A time variable is also needed to distinguish observations for a given unit
3. The observations over several periods for the same unit should be adjacent. This is called long form of panel data

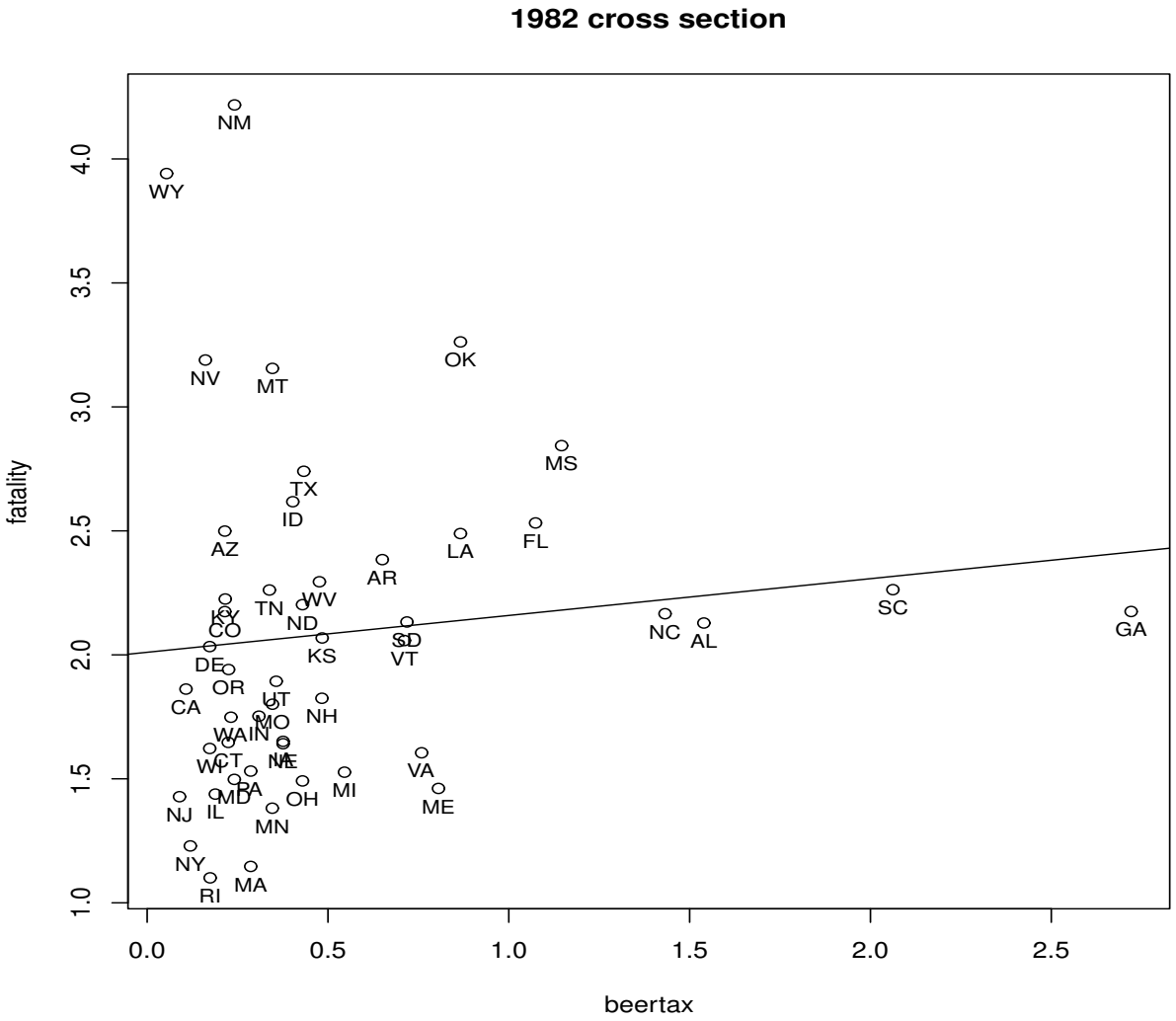
Example: fatality data

```
> ad = "https://www.fsb.miamioh.edu/lij14/411_fatality.txt"
> data = read.table(url(ad), header=T)
> head(data)
  statename year unrate  beertax  mlda     pop  pop1517 unus       y
1        AL 1982   14.4 1.539379 19.00 3942002 208999.6  9.7 2.12836
2        AL 1983   13.7 1.788991 19.00 3960008 202000.1  9.6 2.34848
...
6        AL 1987    7.8 1.560000 21.00 4082999 204999.8  6.2 2.71859
> attach(data)
> data[order(year,statename),][1:3,]
   statename year unrate   beertax mlda      pop   pop1517 unus       y
1         AL 1982   14.4 1.5393795   19  3942002  208999.6  9.7 2.12836
15        AR 1982    9.8 0.6503580   21  2306998  122000.0  9.7 2.38405
8         AZ 1982    9.9 0.2147971   19  2896996  141000.0  9.7 2.49914
...
```

Remarks

1. We use Fatality data, and we are interested in whether rising beer tax rate can reduce traffic death. The dependent variable y is traffic death, and the key regressor is beer tax rate
2. We observe each state year after year, from 1982 to 1988. So this is panel data
3. ID variable is statename; time variable is year
4. Each state is a unit or cluster. Here each state has 7 yearly observations
5. In short, observation is at state-year level
6. The last code sorts the panel data to show the 1982 cross section, that is, all states in 1982

1982 Cross Section



Remarks

1. We need to understand the flaw of cross section data before discussing the advantage of panel data
2. For instance, the last graph displays the 1982 cross section. The fitted OLS line is upward-sloping. Taken literally, it means increasing beer tax would increase traffic death
3. This counter-intuitive finding reflects bias of omitted variables such as weather, speed limit, drinking culture, etc
4. For instance, it is problematic to compare Wyoming (apple) to Ohio (orange). Nevertheless, the cross section regression is based on that between-state variation

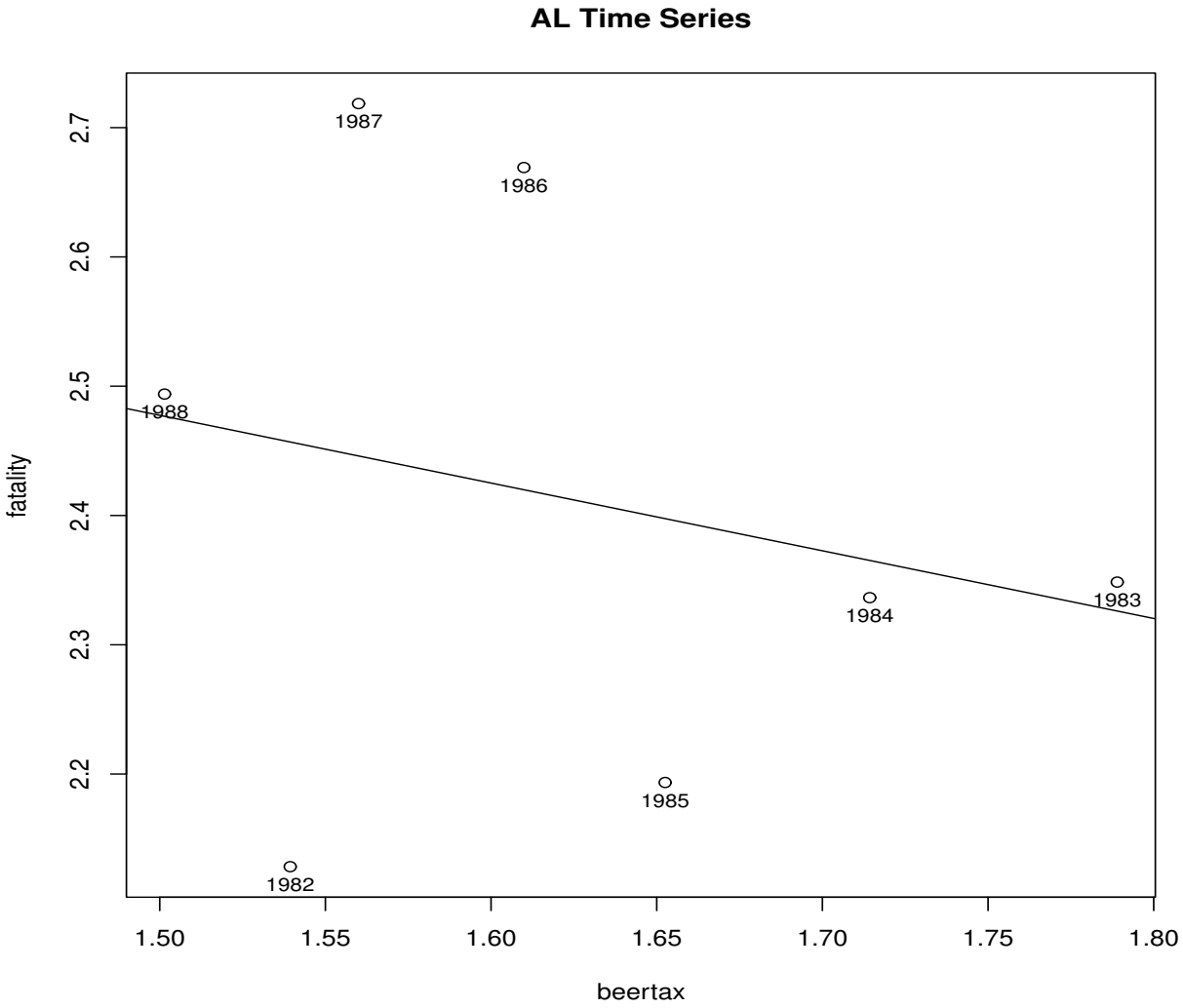
(Bad) OLS Regression with Cross Section Data

```
> summary(lm(y~beertax, subset=(year==1982)))$coef
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 2.0103813  0.1390785 14.4550158 1.010417e-18
beertax     0.1484603  0.1883681  0.7881393 4.346580e-01

> summary(lm(y~beertax))$coef
             Estimate Std. Error  t value      Pr(>|t|)
(Intercept) 1.8533067 0.04356724 42.53900 6.951295e-137
beertax     0.3646064 0.06216998  5.86467  1.082160e-08
```

The estimated coefficients of beer tax are positive, no matter just one cross section or all cross sections are used. Those two regressions are prone to substantial omitted variable bias!

Alabama Time Series



Remarks

1. A better approach is looking at data from a different angle, and focusing on the same state over time or time series data
2. By doing so, we can hold constant confounders that are time-invariant for given state
3. For instance, we can hold speed limit constant if it does not change over time for given state
4. As an example, the last graph shows that OLS line becomes downward-sloping if the time series data of Alabama are used to fit regression

(Good) OLS Regression with Time Series Data

```
> summary(lm(y~beertax, subset=(statename=="AL")))$coef
              Estimate Std. Error    t value   Pr(>|t|)
(Intercept)  3.2631384  1.5583186  2.0940124 0.09043795
beertax     -0.5237807  0.9580533 -0.5467135 0.60807890
```

1. The estimated coefficient of beer tax is negative—we are on the right track now
2. That improved result is due to the fact that a time series regression uses the constant term as proxy for time-invariant unobserved factors such as location and speed limit of a state
3. Another way to explain the improvement is that the time series regression makes use of within-state variation
4. The time series regression suffers a small sample size, and can be improved further by panel data regression

(Better) One Way Fixed Effect Regression with Panel Data

The idea of using a constant term to proxy time-invariant confounders can be generalized—we need to include a set of dummy variables, one for each state. Those dummy variables can serve as proxy for confounders that are time-invariant for a given state, but vary across states (eg., WY and OH may have different speed limits. But within the state the speed limit does not change over time)

$$y_{it} = \beta x_{it} + \sum_{i \neq 1} \alpha_i D_i + u_{it} \quad (\texttt{One Way Fixed Effect Model}) \tag{1}$$

1. i is state index, and t is time index
2. Dummy variable D_i equals one for the i-th state. We call them state fixed effect
3. We drop D_1 to avoid dummy variable trap
4. Omitted variable bias arises if significant state fixed effect is excluded from regression

An Equivalent Form

Let's look at $\sum_i \alpha_i D_i$ more closely. Suppose each state has three yearly observations

$$\alpha_{AL} D_{AL} + \alpha_{AR} D_{AR} + ... = \alpha_{AL} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix} + \alpha_{AR} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix} + ... = \begin{pmatrix} \alpha_{AL} \\ \alpha_{AL} \\ \alpha_{AL} \\ \alpha_{AR} \\ \alpha_{AR} \\ \alpha_{AR} \\ \vdots \end{pmatrix} \tag{2}$$

That is why model (1) can be rewritten as

$$y_{it} = \beta x_{it} + \alpha_i + u_{it}, \quad (\texttt{one way FE model}) \tag{3}$$

Note that the state fixed effects α_i are actually coefficients of state dummy variables

State Fixed Effect

$$\begin{pmatrix} \alpha_{AL} \\ \alpha_{AL} \\ \alpha_{AL} \\ \alpha_{AR} \\ \alpha_{AR} \\ \alpha_{AR} \\ \vdots \end{pmatrix} = \begin{pmatrix} \texttt{AL speed limit} + \texttt{AL terrain} + \texttt{AL weather} + \ldots \\ \texttt{AL speed limit} + \texttt{AL terrain} + \texttt{AL weather} + \ldots \\ \texttt{AL speed limit} + \texttt{AL terrain} + \texttt{AL weather} + \ldots \\ \texttt{AR speed limit} + \texttt{AR terrain} + \texttt{AR weather} + \ldots \\ \texttt{AR speed limit} + \texttt{AR terrain} + \texttt{AR weather} + \ldots \\ \texttt{AR speed limit} + \texttt{AR terrain} + \texttt{AR weather} + \ldots \\ \vdots \end{pmatrix} \tag{4}$$

Thus, all confounders that are time-variant (such as speed limit, terrain, weather) are accounted for by the state dummy variables. They no longer cause trouble (bias) once state dummy variables are included in the regression!

One Way Fixed Effect Regression

```
 > summary(lm(y~beertax+factor(statename)))$coef
                          Estimate Std. Error    t value     Pr(>|t|)
(Intercept)           3.47763139  0.3133586 11.0979282 4.754084e-24
beertax              -0.65587452  0.1878511 -3.4914598 5.560018e-04
factor(statename)AR -0.65495241  0.2190216 -2.9903550 3.028110e-03
...
factor(statename)WY -0.22850500  0.3128977 -0.7302866 4.658107e-01
```

1. R factor function can be used as a shortcut for those state dummy variables. The results are long because there are so many dummy variables. So here I only show some results
2. AL is base group, and its dummy variable is absent
3. Good news: the coefficient of beer tax is -0.65587452, a negative number, and its t value -3.4914598 exceeds 1.96 in absolute value
4. The t value for the AR dummy variable is -2.9903550, implying that there is a significant difference between AR and AL. In other words, omitted variable bias would arise if we do not include the AR dummy variable

Statistical Significance of State Fixed Effect

```
> femodel = lm(y~beertax+factor(statename))
> poololsmodel = lm(y~beertax)
> anova(poololsmodel,femodel)
Analysis of Variance Table
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    334 98.747
2    287 10.345 47    88.402 52.179 < 2.2e-16 ***
```

1. We can test the statistical significance of state fixed effect. The null hypothesis is that those state dummy variables all have zero coefficients (no state fixed effect)

2. Recall that F test is synonymous with Analysis of Variance (ANOVA). In this case, the unrestricted model is the regression with state fixed effect, while the restricted model is the regression without state fixed effect

3. F value is 52.179, and its p value is less than 0.05. So the null hypothesis is rejected. This finding explains why pooled OLS is biased—it omites the state fixed effect!

Economic Significance of State Fixed Effect

The economic significance of state fixed effect can be evaluated by comparing the R-squared of regressions with and without state fixed effect

```
> summary(lm(y~beertax))$r.squared
[1] 0.09336287
> summary(lm(y~beertax+factor(statename)))$r.squared
[1] 0.905014
```

For fatality data, including state fixed effect increases the R squared from 0.09336287 to 0.905014, a tremendous improvement! Thus, even if the goal is forecast other than causality, using fixed effect can substantially enhance the fit of model

(Even Better) Two Way Fixed Effect Regression with Panel Data

We can do even better by including time dummy variables, one for each period. Those dummy variables can serve as proxy for confounders that vary over time, but is the same for all states at a given period (eg., cars become safer over time, and that happens in all states)

$$y_{it} = \beta x_{it} + \sum_{i \neq 1} \alpha_i D_i + \sum_{t \neq 1} \gamma_t D_t + u_{it} \quad (\texttt{Two Way Fixed Effect Model}) \tag{5}$$

1. Dummy variable D_t equals one for the t-th period. We call them time fixed effect
2. We drop the first-period dummy variable to avoid dummy variable trap
3. Omitted variable bias arises if significant time fixed effect are excluded from regression

Regression (5) can be rewritten as

$$y_{it} = \beta x_{it} + \alpha_i + \gamma_t + u_{it} \quad (\texttt{Two Way Fixed Effect Model}) \tag{6}$$

Two Way Fixed Effect Regression

```
> summary(lm(y~beertax+factor(statename)+factor(year)))$coef
                          Estimate Std. Error    t value     Pr(>|t|)
(Intercept)           3.51137066 0.33250321 10.5604112 3.564251e-22
beertax              -0.63997707 0.19737793 -3.2423943 1.328188e-03
factor(statename)AR -0.63852689 0.22732144 -2.8089162 5.319549e-03
...
factor(statename)WY -0.20345658 0.32680702 -0.6225588 5.340794e-01
factor(year)1983     -0.07990287 0.03835395 -2.0833021 3.812752e-02
...
factor(year)1988     -0.05180363 0.03962371 -1.3073899 1.921491e-01
```

1. AL and 1982 are base groups, and their dummy variables are absent
2. Good news: the coefficient of beer tax -0.63997707 is still negative and significant
3. The t value for the 1983 dummy variable is -2.0833021, implying that there is a significant difference between 1983 and 1982

Adding Covariates

```
> data$youngratio = pop1517/pop
> summary(lm(y~beertax+youngratio+mlda+factor(statename)+factor(year),data
                          Estimate Std. Error    t value     Pr(>|t|)
(Intercept)            3.26927968 0.73659084  4.4383931 1.306442e-05
beertax               -0.64304568 0.19783211 -3.2504616 1.293451e-03
youngratio            -2.15970834 8.82492901 -0.2447281 8.068468e-01
mlda                   0.01859684 0.02141223  0.8685149 3.858590e-01
factor(statename)AR -0.66310063 0.22936264 -2.8910577 4.141399e-03
...
```

1. State fixed effect and time fixed effects (those dummy variables) can account for confounders that are fixed either in i dimension or t dimension
2. For confounders that vary in both i dimension and t dimension, we need to try our best to find data and control them. Ignoring them will lead to OVB!
3. For example, ratio of young drivers with age between 15-17 and minimum legal driving age are controlled for here. The coefficient of beer tax remains negative and significant!

Panel Data Is Very Useful for Causal Study

1. The unit fixed effect α_i and time fixed effect γ_t can take care of a lot of unobserved factors
2. That is why fixed effect regression with panel data can suffer much less bias than regression with cross section data. That is also why panel data is widely used for causal study
3. Bias can still be present for fixed effect regression if $cov(u_{it}, x_{it}) \neq 0$. To correct that kind of endogeneity, we need other weapons such as DID and instrumental variable

Within Regression

1. Recall that if we regress y onto a constant term $y = \beta_0 + u$, The estimate is the sample average $\hat{\beta}_0 = \bar{y}$, and residual is $y - \bar{y}$

2. According to FW theorem, β in the one way fixed regression can be obtained based on residuals of regressing y and x onto state dummy variables, and those residuals look like

$$\tilde{y}_{it} = y_{it} - \bar{y}_i, \quad \bar{y}_i = \frac{\sum_{t=1}^{T} y_{it}}{T}, \quad (\texttt{within transformation of y}) \tag{7}$$

$$\tilde{x}_{it} = x_{it} - \bar{x}_i, \quad \bar{x}_i = \frac{\sum_{t=1}^{T} x_{it}}{T}, \quad (\texttt{within transformation of x}) \tag{8}$$

3. Within regression uses $\tilde{y}_{it}$ as the dependent variable, and $\tilde{x}_{it}$ as the independent variable. Because within regression is the step II of FW theorem, it produces the same β estimate as the fixed effect regression

4. The within transformation can remove α_i since $\tilde{\alpha}_i = \alpha_i - \bar{\alpha}_i = 0$. That is another way to explain why fixed effect regression is good for causal study—it removes troublemakers!

Within Regression in R

```
> library(plm)
> data_panel = pdata.frame(data, index = c("statename", "year"))
> summary(plm(y~beertax, data = data_panel, model = "within"))$coef
          Estimate Std. Error  t-value     Pr(>|t|)
beertax -0.6558745  0.1878511 -3.49146 0.0005560018
```

1. To run within regression in R, we need to (i) define a panel data frame using pdata.frame function in the plm package, and (ii) use plm function with model="within" option
2. The β estimate and its t value are the same as the one way fixed effect regression

Serial Correlation and Standard Error

Suppose there is a sample of two time series observations (y_1, y_2) with equal variance $var(y_1) = var(y_2) = \sigma^2$. The data are i.d, not i.i.d due to serial correlation

$$cov(y_1, y_2) \neq 0, \quad (\texttt{serial correlation}) \tag{9}$$

Then we can show the standard error of sample mean is

$$se(\bar{y}) = \sqrt{var(\bar{y})} = \sqrt{var\left(\frac{y_1 + y_2}{2}\right)} = \sqrt{\frac{2\sigma^2 + 2cov(y_1, y_2)}{4}} \neq \frac{\sigma}{\sqrt{2}} \tag{10}$$

1. $\frac{\sigma}{\sqrt{2}}$ is the conventional standard error, which is valid only if data are independent $cov(y_1, y_2) = 0$
2. $\sqrt{\frac{2\sigma^2 + 2cov(y_1, y_2)}{4}}$ is robust standard error, which is always valid no matter data are dependent or independent
3. We need cluster robust standard error to account for serial correlation within a unit

Cluster Robust Standard Error

```
> library(sandwich)
> library(lmtest)
> m = plm(y~beertax, data = data_panel, model = "within")
> cluster_robust_se = vcovHC(m, type = "HC1", cluster = "group")
> coeftest(m, vcov. = cluster_robust_se)
        Estimate Std. Error t value Pr(>|t|)
beertax -0.65587    0.28880  -2.271  0.02388
```

1. In general, $cov(y_1, y_2) > 0$. So the robust se is greater than conventional se
2. That is the case here: the cluster robust se 0.28880 is greater than the conventional se 0.1878511
3. The absolute robust t value 2.271 is less than conventional t value 3.49146, accordingly

Summary

With panel data, a baseline model is given as

$$y_{it} = \beta x_{it} + \theta_{it} + \alpha_i + \gamma_t + u_{it} \quad \texttt{(Two Way Fixed Effect Model with Covariates)} \tag{11}$$

1. x_{it} is the variable of interest. It has to vary over time
2. θ_{it} is time-varying control variable or covariate; α_i is unit fixed effect; γ_t is time fixed effect
3. $\hat{\beta}$ is unbiased if exogeneity condition $cov(x_{it}, u_{it}) = 0$ is satisfied
4. Cluster robust standard error is required to account for $cov(u_{it}, u_{it-j}) \neq 0$