Advanced Data Analysis I

Panel Data Part 2

PA 541 Week 14

Michael D. Siciliano
Department of Public Administration
College of Urban Planning and Public Affairs

Overview

- The slides and script will discuss:
  - Panel data (sometimes called longitudinal data or time-series-cross-section)
  - First Differencing Methods (FD)
  - Fixed Effects Transformation (FE) or the “within” method.
  - Dummy Variable Fixed Effects Regression

Example Datasets used in lecture are from Wooldridge chapters 13 and 14.

In Class Exercise – Pooled Cross-Sectional Analysis

- Using the cps_inclass dataset, build a single regression model to assess:
  - Whether the gender gap in wages has increased or decreased between 1978 and 1985
  - Whether the return to education has changed between 1978 and 1985.
  - Include in your model the other following variables: y85 + exper + expersq + union

- The variables in the cps dataset are as follow:
  - 1. educ years of schooling
  - 2. south =1 if live in south
  - 3. nonwhite =1 if nonwhite
  - 4. female =1 if female
  - 5. married =1 if married
  - 6. exper age - educ - 6
  - 7. expersq exper^2
  - 8. union =1 if belong to union
  - 9. lwage log hourly wage
  - 10. age in years
  - 11. year 78 or 85
  - 12. y85 =1 if year == 85

```r
cps=read_csv(file="cps_inclass.csv")
cps
```

```
# A tibble: 1,084 x 13
      X1  educ south nonwhite female married exper expersq union lwage   age  year   y85
   <dbl> <dbl> <dbl>    <dbl>  <dbl>   <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1     1    12     0        0      0       0     8      64     0 1.22     25    78     0
 2     2    12     0        0      1       1    30     900     1 1.61     47    78     0
 3     3     6     0        0      0       1    38    1444     1 2.14     49    78     0
 4     4    12     0        0      0       1    19     361     1 2.07     36    78     0
 5     5    12     0        0      0       1    11     121     0 1.65     28    78     0
 6     6     8     0        0      0       1    43    1849     0 1.71     56    78     0
 7     7    11     0        0      0       0     2       4     0 1.10     18    78     0
 8     8    15     0        0      1       0     9      81     0 1.83     29    78     0
 9     9    16     0        0      1       0    17     289     0 0.357    38    78     0
10    10    15     0        0      0       1    23     529     1 2.15     43    78     0
# ... with 1,074 more rows
```

```r
#create the interaction terms
cps$y85educ=cps$educ * cps$y85
cps$y85fem=cps$female * cps$y85

cps1b=lm(lwage ~ y85 + educ + exper + expersq + union + female +
           y85fem + y85educ, data=cps)
summary(cps1b)
```

```
> summary(cps1b)

Call:
lm(formula = lwage ~ y85 + educ + exper + expersq + union + female +
    y85fem + y85educ, data = cps)

Residuals:
     Min       1Q   Median       3Q      Max
-2.56098 -0.25828  0.00864  0.26571  2.11669

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.589e-01  9.345e-02   4.911 1.05e-06 ***
y85          1.178e-01  1.238e-01   0.952   0.3415
educ         7.472e-02  6.676e-03  11.192  < 2e-16 ***
exper        2.958e-02  3.567e-03   8.293 3.27e-16 ***
expersq     -3.994e-04  7.754e-05  -5.151 3.08e-07 ***
union        2.021e-01  3.029e-02   6.672 4.03e-11 ***
female      -3.167e-01  3.662e-02  -8.648  < 2e-16 ***
y85fem       8.505e-02  5.131e-02   1.658   0.0977 .
y85educ      1.846e-02  9.354e-03   1.974   0.0487 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4127 on 1075 degrees of freedom
Multiple R-squared:  0.4262,    Adjusted R-squared:  0.4219
F-statistic:  99.8 on 8 and 1075 DF,  p-value: < 2.2e-16
```

Panel data

$$Crime_{it} = \beta_0 + \beta_1 Police_{i,t-1} + \epsilon_{it} \tag{8.1}$$

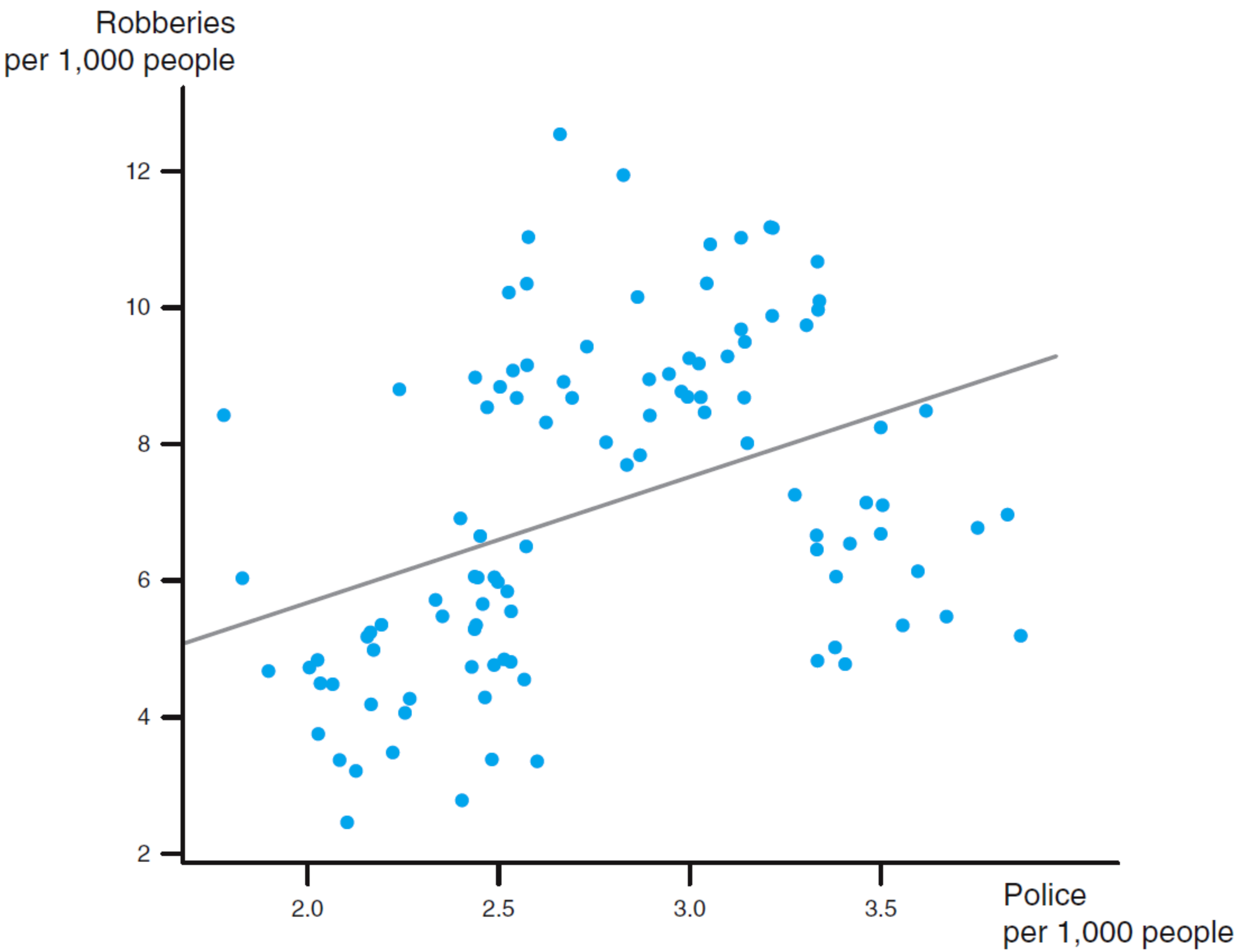


Image from Bailey (2016)

FIGURE 8.1: Robberies and Police for Large Cities in California

$$Crime_{it} = \beta_0 + \beta_1 Police_{i,t-1} + \epsilon_{it} \quad (8.1$$

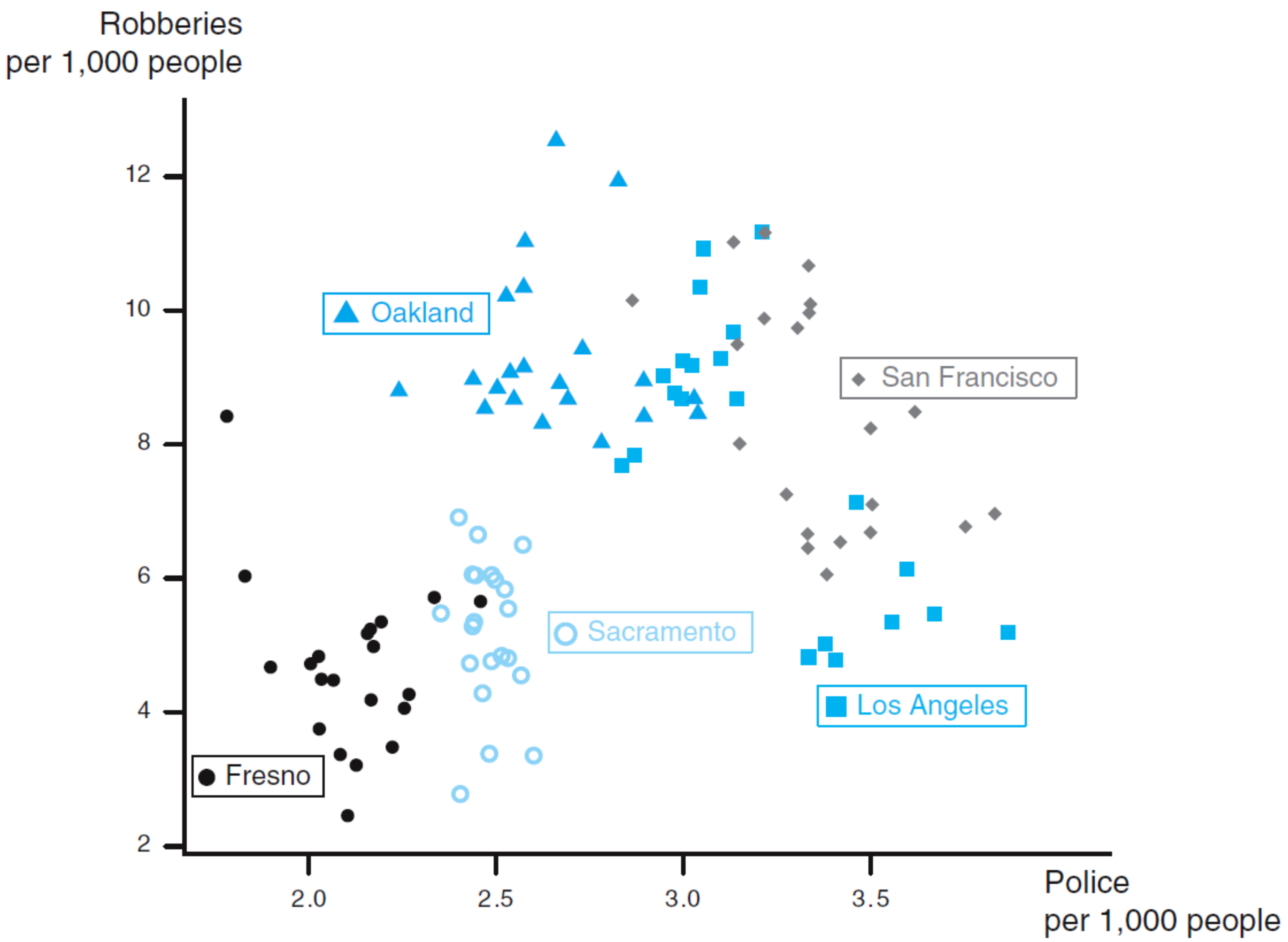


Image from Bailey (2016)

FIGURE 8.2: Robberies and Police for Specified Cities in California

$$Crime_{it} = \beta_0 + \beta_1 Police_{i,t-1} + \epsilon_{it} \quad (8.1)$$

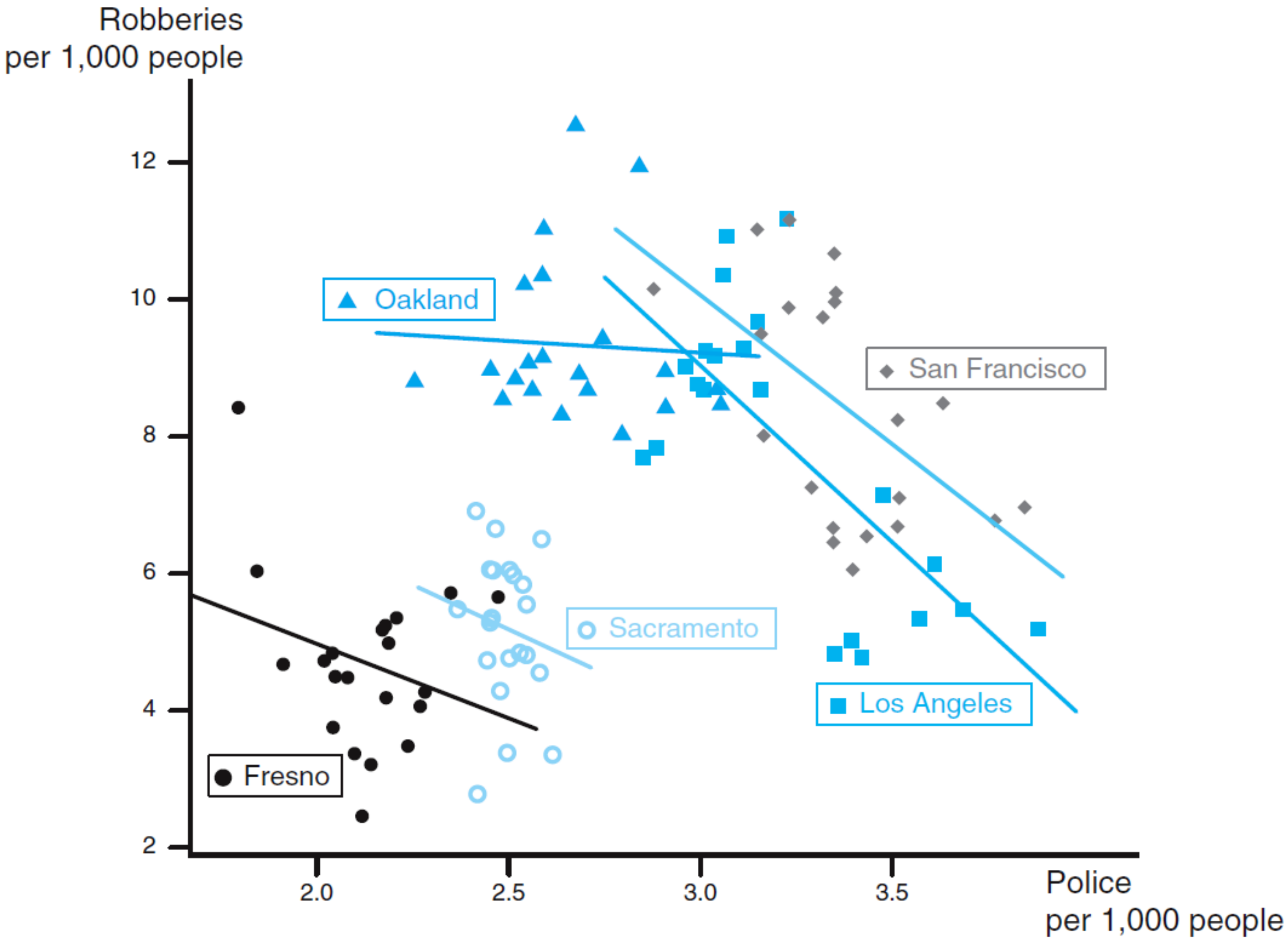
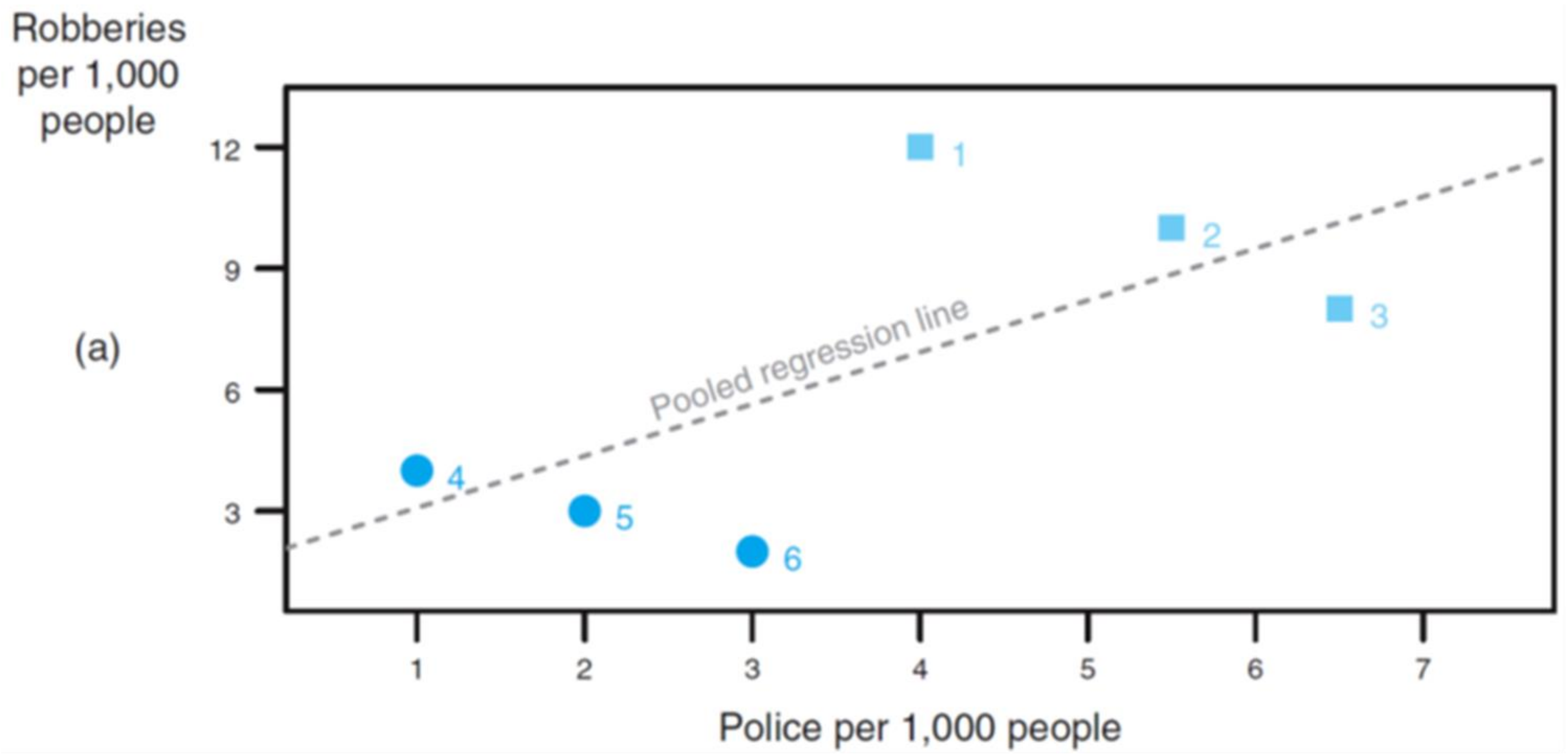


FIGURE 8.3: Robberies and Police for Specified Cities in California with City-Specific Regression Lines

Image from Bailey (2016)

- In a pooled model a common source of endogeneity is that the specific units (here cities) have different baseline levels of crime and these levels are correlated with our independent variables. Thus, cities with higher crime also tend to have more police. This creates a positive correlation in the pooled model.



Example: Test scores

$$Test\ scores_{it} = \beta_0 + \beta_1 Private\ school_{it} + \epsilon_{it} \qquad (8.2)$$

1) What is in the error term?

2) Are there any stable unit-specific elements in the error term?

3) Are the stable unit-specific element in the error term correlated with the independent variable?

Pooled Cross Section versus Panel Data

- **Pooled Cross Section Data**
  - Random samples taken at different time points…thus pulling produces one large random sample with a time element.
  - Time dummy variables can be used to capture structural change over time and can be interacted with other variables.
  - Observations across different time periods provide an opportunity to evaluate the effect of policies or programs.

- **Panel Data**
  - Unlike pooled cross-sectional data, panel data contains cross sections of the same individuals/units at different points in time.
  - This structure allows us to analyze more complicated models that we may believe more accurately reflect the real world.
    - The main benefit, as we will see, is our ability to account for unobserved heterogeneity.

Structure of Panel Data

Panel data can be in two formats long or wide. Wide data stores each variable separately for each wave, so only has one observation for each individual:

PID	inc1	inc2	inc3	inc4
1	200	210	220	250
2	600	660	700	750
3	250	280	200	210
4	150	190	250	300

Long data stores all observations of a variable, for example income, in the same variable, and has a wave variable and multiple observations for each individual:

PID	wave	inc
1	1	200
1	2	210
1	3	220
1	4	250
2	1	600
2	2	660
2	3	700
2	4	750

For a quick tutorial on how to do this in R using the new pivot_wider and pivot_longer functions go to: https://tidyr.tidyverse.org/articles/pivot.html
I also have examples in the script we can look at.

Types of Panel Data Methods

- Fixed effects estimation
    1. **First Difference methods**
    2. **Fixed effects estimation** or within transformation
    3. **Dummy variable fixed effects model**

- Longitudinal Multilevel Models/Random effects estimation are not covered in this course.

1. Panel Data and First Differencing

Two period panel data analysis - Unobserved effects model

- What is the effect of unemployment on crime?
- We have a dataset with two years of data, 1982 and 1987, that contains crime rate and unemployment rate information for 46 cities.
- First, let's look at the effect of unemployment on crime from the 1987 cross section.

```r
#let's look at the relationship in just 1987
crime2=filter(crime, year==87)#subset the data into a new file with just 1987 data
crime2
crm1=lm(crmrte ~ unem, data=crime2)
summary(crm1)
```

```
> crime2=subset(crime, year==87)#subset the data into a new file with just 19
87 data
> crm1=lm(crmrte ~ unem, data=crime2)
> summary(crm1)

Call:
lm(formula = crmrte ~ unem, data = crime2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  128.378     20.757   6.185  1.8e-07 ***
unem          -4.161      3.416  -1.218     0.23    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.6 on 44 degrees of freedom
Multiple R-squared:  0.03262,	Adjusted R-squared:  0.01063 
F-statistic: 1.483 on 1 and 44 DF,  p-value: 0.2297
```

- If we interpret this causally, it implies an increase in unemployment rate lowers the crime rate. The coefficient is not statistically significant…we have, at best, found no link between crime and unemployment.

Note this is just 1987 data

```r
#Let's now simply pool the data together and use standard OLS
crime$d87.2=ifelse (crime$year==87,1,0)#though it was already made in the dataset, here
#is how I would create a simple year dummy variable

crm2=lm(crmrte ~ d87 + unem, data=crime)
summary(crm2)
```

```
Call:
lm(formula = crmrte ~ d87 + unem, data = crime)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.474 -21.794  -6.266  18.297  75.113 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  93.4203    12.7395   7.333 9.92e-11 ***
d87           7.9404     7.9753   0.996    0.322    
unem          0.4265     1.1883   0.359    0.720    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.99 on 89 degrees of freedom
Multiple R-squared:  0.01221,	Adjusted R-squared:  -0.009986 
F-statistic: 0.5501 on 2 and 89 DF,  p-value: 0.5788
```

- An alternative way to use panel data is to view the unobserved factors affecting the dependent variable as consisting of two types: those that are constant and those that vary over time.
- Letting *i* denote the cross-sectional unit and *t* the time period, we can write:

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}, \quad t = 1,2$$

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}, \quad t = 1,2$$

- In the notation, *i* denotes the person, firm, city, etc… and *t* denotes the time period.
- The variable $d2_t$ is a dummy variable which equals zero when t=1 and one when t = 2; it does not change across *i*, which is why it has no *i* subscript.
- As you can see the intercept is allowed to vary overtime.
- The variable a_i captures all unobserved time constant factors that affect y_{it}. This is referred to as an unobserved effect or fixed effect. You might also see a_i referred to as unobserved heterogeneity.

An unobserved effects model

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + u_{it}$$

- In this model, a_i captures all factors affecting city crime rates that do not change overtime. We can think of it as the unobserved city effect; something about the city that is a fixed difference from other cities (location, transit, education level, demographics…while some of these may change, they may be roughly constant over the observed time period).
- How should we estimate the effect of unemployment given our two years of panel data?

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + u_{it}$$

- One approach, as we just saw, is to just pool the data together and use OLS.
  - In order to produce a consistent estimator we must assume the unobserved effect, a_i is uncorrelated with x_{it}. Note that since a_i is not measured, it becomes part of the error term – this is often referred to as a composite error.

- As long as a_i is correlated with our regressor, we have **unobserved heterogeneity** in our model.
- In our crime example, we believe that the unmeasured city factors that influence the crime rate are correlated with the unemployment rate.
- The availability of observations of the same individuals as multiple time points allows us to overcome this heterogeneity bias. This is a key reason for collecting and analyzing panel data.

- Because a_i is constant overtime, we can difference the data across the two years.

Analyzing our data with pooled OLS we have:

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}, \; t = 1{,}2$$

If we look at the equation for each time period (i.e. when t=2 and when t =1):

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2}, \; t = 2$$

$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1}, \; t = 1$$

If we subtract the second equation from the first

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

Which can be rewritten as

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

- This is the first differenced equation. It is just a cross-sectional equation, but each variable is differenced overtime.
- Note, that a_i is no longer in the equation.

First differenced results

- We now find a positive and statistically significant effect on unemployment.
- What does the intercept tell us?

The costs of differencing our data

- Though the first differences approach allows us to control for unobserved effects, it can also:
  - Reduce the variation in the explanatory variables as we are only looking at changes within each unit of observation. Variables such as education which change very infrequently (if at all) for adult full-time workers are difficult to estimate with such a model since there is little variation in Δx_i.
  - Any predictors, such as race or gender, which do not change overtime are dropped from the model as the difference score for each person is 0 and thus there is no variation at all.

2. Panel data and Fixed effects estimation

Fixed effects estimation

- We saw that first-differencing yields unbiased parameter estimates by eliminating the time-constant fixed effect, $a_{i.}$
- However, there is an alternative method known as the fixed effects transformation that can be used to estimate the model (and this tends to be much more commonly seen in published research).
- In this method we time demean the data on y_{it}, x_{it}, and the error term.
  - This approach is also referred to as the within effect estimator.

Fixed effects estimation

- For the fixed effects transformation we begin with a single explanatory variable for each observational unit *i*:

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \quad t = 1,2, \ldots . T$$

- For each *i* we average this equation over time:

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i$$

- Then subtract the second equation from the first:

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i , \quad t = 1,2, \ldots . T$$

Or

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it} , \quad t = 1,2, \ldots . T$$

Note, since we can think of a_i as individual intercept terms, we do not need to have an overall intercept term.

Fixed effects transformation

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it}, \quad t = 1,2, \ldots .T$$

- Here we have time demeaned data. Notice, as with first differencing, the a_i is removed.
- A general time-demeaned equation for each i will be:

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{it2} + \cdots + \beta_k \ddot{x}_{itk} + \ddot{u}_{it}, \quad t = 1,2, \ldots .T$$

- This estimator is also known as the ***within*** estimator, because all of the variation in the dependent and independent variables are ***within*** each cross-sectional unit.

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{it2} + \cdots + \beta_k \ddot{x}_{itk} + \ddot{u}_{it}, \; t = 1,2,\ldots.T$$

- The primary assumption remains that the error term is uncorrelated with the regressors.

- As with the first differenced estimator, the coefficients on any regressor that remains constant over time cannot be estimated.

- Applying OLS to this model leads to unbiased parameter estimates – though we need to adjust our degrees of freedom to get correct standard errors. For this reason, and to allow the software to time demean the data for us, fixed effects regression models are available in most statistical packages.

Panel data methods in R

- Need the plm package.
- As with Stata, you need to define the cross-sectional and time elements of our data. We do this with the pdata.frame() function.
  - p.mom = pdata.frame(mom, index=c("id", "year"))
  - Here, we are telling R that the dataset ‘mom’ should be stored as a panel data frame and that the group and time variables are called ‘id’ and ‘year’ respectively.
- To model either FE or FD we use the function plm () and indicate model = “within” or model = “fd”.
  - mod3 = plm(lnhr ~ lnwg, data=p.mom, model="within")
- Let’s look at a dataset and run a fixed effect model.

```
> head(nccrime)
  county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc  density    taxpc
1      1   81 0.0398849 0.289696 0.402062 0.472222   5.61 0.00178678 2.307159 25.69763
2      1   82 0.0383449 0.338111 0.433005 0.506993   5.59 0.00176659 2.330254 24.87425
3      1   83 0.0303048 0.330449 0.525703 0.479705   5.80 0.00183577 2.341801 26.45144
4      1   84 0.0347259 0.362525 0.604706 0.520104   6.89 0.00188588 2.346420 26.84235
5      1   85 0.0365730 0.325395 0.578723 0.497059   6.55 0.00192436 2.364896 28.14034
6      1   86 0.0347524 0.326062 0.512324 0.439863   6.90 0.00189522 2.385681 29.74098
  west central urban pctmin80     wcon      wtuc     wtrd     wfir     wser   wmfg
1    0       1     0  20.2187 206.4803  333.6209 182.3330 272.4492 215.7335 229.12
2    0       1     0  20.2187 212.7542  369.2964 189.5414 300.8788 231.5767 240.33
3    0       1     0  20.2187 219.7802 1394.8035 196.6395 309.9696 240.1568 269.70
4    0       1     0  20.2187 223.4238  398.8604 200.5629 350.0863 252.4477 281.74
5    0       1     0  20.2187 243.7562  358.7830 206.8827 383.0707 261.0861 298.88
6    0       1     0  20.2187 257.9139  369.5465 218.5165 409.8842 269.6129 322.65
    wfed   wsta   wloc        mix    pctymle d82 d83 d84 d85 d86 d87    lcrmrte    lprbarr
1 409.37 236.24 231.47 0.09991788 0.08769682   0   0   0   0   0   0 -3.221757 -1.238923
2 419.70 253.88 236.79 0.10304912 0.08637666   1   0   0   0   0   0 -3.261134 -1.084381
3 438.85 250.36 248.58 0.08067867 0.08509085   0   1   0   0   0   0 -3.496449 -1.107303
4 459.17 261.93 264.38 0.07850353 0.08383328   0   0   1   0   0   0 -3.360270 -1.014662
5 490.43 281.44 288.58 0.09324856 0.08230646   0   0   0   1   0   0 -3.308445 -1.122715
6 478.67 286.91 306.70 0.09732283 0.08008062   0   0   0   0   1   0 -3.359507 -1.120668
    lprbconv   lprbpris  lavgsen    lpolpc  ldensity   ltaxpc    lwcon    lwtuc    lwtrd
1 -0.9111490 -0.7503061 1.724551 -6.327340 0.8360171 3.246399 5.330205 5.810005 5.205835
2 -0.8370060 -0.6792581 1.720979 -6.338704 0.8459773 3.213833 5.360137 5.911600 5.244607
3 -0.6430188 -0.7345839 1.757858 -6.300291 0.8509204 3.275311 5.392628 7.240509 5.281372
4 -0.5030129 -0.6537265 1.930071 -6.273361 0.8528909 3.289981 5.409070 5.988612 5.301128
5 -0.5469313 -0.6990466 1.879465 -6.253162 0.8607340 3.337204 5.496169 5.882718 5.332152
6 -0.6687981 -0.8212920 1.931521 -6.268420 0.8694848 3.392526 5.552626 5.912277 5.386862
     lwfir    lwser    lwmfg    lwfed    lwsta    lwloc      lmix  lpctymle  lpctmin
1 5.607452 5.374044 5.434246 6.014619 5.464848 5.444450 -2.303407 -2.433870 3.006608
2 5.706707 5.444911 5.482013 6.039540 5.536862 5.467174 -2.272549 -2.449038 3.006608
3 5.736475 5.481292 5.597310 6.084157 5.522900 5.515765 -2.517281 -2.464036 3.006608
4 5.858180 5.531204 5.640985 6.129421 5.568077 5.577387 -2.544612 -2.478925 3.006608
5 5.948220 5.564850 5.700042 6.195282 5.639919 5.664972 -2.372487 -2.497306 3.006608
6 6.015875 5.596987 5.776568 6.171011 5.659169 5.725870 -2.329722 -2.524721 3.006608
     clcrmrte     clprbarr    clprbcon    clprbpri     clavgsen     clpolpc     cltaxpc
1          NA           NA          NA          NA           NA          NA          NA
2 -0.03937626  0.154542208  0.07414299  0.07104796 -0.003571391 -0.01136398 -0.03256536
3 -0.23531556 -0.022922039  0.19398713 -0.05532581  0.036878586  0.03841305  0.06147742
4  0.13617969  0.092641115  0.14000595  0.08085740  0.172213197  0.02693033  0.01467013
5  0.05182457 -0.108053565 -0.04391843 -0.04532003 -0.050606012  0.02019882  0.04722309
6 -0.05106163  0.002047777 -0.12186676 -0.12224543  0.052056313 -0.01525831  0.05532193
```

Crime related data for 90 counties in North Carolina

Apply the Fixed Effects Transformation to our North Carolina Crime Data

```
> plm.nc=pdata.frame(nccrime, index=c("county","year"))
>
> fem1b=plm(lcrmrte ~ d82 + d83 + d84 + d85 + d86+ d87 + lprbarr + lprbconv +
+              lprbpris + lavgsen + lpolpc, data=plm.nc, model="within")
> summary(fem1b)
Oneway (individual) effect Within Model

Call:
plm(formula = lcrmrte ~ d82 + d83 + d84 + d85 + d86 + d87 + lprbarr +
    lprbconv + lprbpris + lavgsen + lpolpc, data = plm.nc, model = "within")

Balanced Panel: n=90, T=7, N=630


Coefficients :
           Estimate Std. Error  t-value  Pr(>|t|)
d82       0.0125802  0.0215416   0.5840 0.5594712
d83      -0.0792813  0.0213399  -3.7152 0.0002247 ***
d84      -0.1177281  0.0216145  -5.4467 7.871e-08 ***
d85      -0.1119561  0.0218459  -5.1248 4.182e-07 ***
d86      -0.0818268  0.0214266  -3.8189 0.0001499 ***
d87      -0.0404704  0.0210392  -1.9236 0.0549446 .
lprbarr  -0.3597944  0.0324192 -11.0982 < 2.2e-16 ***
lprbconv -0.2858733  0.0212173 -13.4736 < 2.2e-16 ***
lprbpris -0.1827812  0.0324611  -5.6308 2.916e-08 ***
lavgsen  -0.0044879  0.0264471  -0.1697 0.8653154
lpolpc    0.4241142  0.0263661  16.0856 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    17.991
Residual Sum of Squares: 10.179
R-Squared      :  0.43424
      Adj. R-Squared :  0.36462
F-statistic: 36.9107 on 11 and 529 DF, p-value: < 2.22e-16
```

We likely have concerns with our standard errors given the clustering in the data.

Cluster-Robust Standard Errors

- Clustered errors occur for several reasons.
  - You may have multilevel structures based on a stratified sampling design (e.g., kids in schools).
  - Another is panel data; model errors in panel data may be correlated across different time periods for a given individual.
- When clustered errors are not controlled for, and we instead rely on the "default" settings, we often obtain misleadingly small standard errors, large t-stats, and low p-values (Cameron and Miller 2013).
  - One way to control for this is to estimate the regression model and then post-estimation obtain "cluster-robust" standard errors.

Cluster-Robust for OLS models using the lm function in R

- There is a new package called ‘multiway’ that allows you to control for clustering in your data.
- As mentioned above, you estimate your regression as usual, and then adjust the standard errors afterward. This is the same approach we used to deal with heteroskedasticity.
- As we will see below, to implement we:
  - m1.vcovCL = cluster.vcov(mod1, mom$id)
  - coeftest(mod1, m1.vcovCL)

‘id’ is the variable in the dataset ‘mom’ that we want to cluster on. Mod1 is the model for which we want to adjust the standard errors.

Cluster-Robust in plm package

- The multiway package accepts lm and glm objects. Models run with the plm package are of class ‘plm’ and ‘panelmodel’ and cannot be passed to the multiway function.
- This is not a problem as the plm package has its own function to deal with clustering.
  - coeftest(mod1b,vcov=vcovHC(mod1b, cluster="group"))

Similar to the multiway package, we are updating the var-cov matrix directly in the coeftest() call and we specify the clustering as “group”. Recall that in plm we created a pdata.frame that indicated what our grouping variable was (in this case ‘id’).

Let’s use lm and plm to replicate some results

- Cameron and Trivedi (2005, p. 708-710) run several different models exploring the relationship between hours and wages.
- They were interested in the responsiveness of labor supply to changes in wages. They note that the standard textbook model of labor supply suggests that for people already working the effect of a wage increase on labor supply is ambiguous, with an income effect pushing in the direction of less work offsetting a substitution effect in the direction of more work.
- Cross-section analysis of adult males finds a relatively small positive response.
  - However, it is possible that this association is spurious.
  - Panel data can help identify the relationship.

The data

```
> head(mom,20)
   lnhr lnwg kids ageh agesq disab id year
1  7.58 1.91    2   27   729     0  1 1979
2  7.75 1.89    2   28   784     0  1 1980
3  7.65 1.91    2   29   841     0  1 1981
4  7.47 1.89    2   30   900     0  1 1982
5  7.50 1.94    2   31   961     0  1 1983
6  7.50 1.93    2   32  1024     0  1 1984
7  7.56 2.12    2   33  1089     0  1 1985
8  7.76 1.94    2   34  1156     0  1 1986
9  7.86 1.99    2   35  1225     0  1 1987
10 7.82 1.98    2   36  1296     0  1 1988
11 7.20 2.54    4   35  1225     0  2 1979
12 6.95 2.52    3   37  1369     1  2 1980
13 7.24 2.59    3   37  1369     1  2 1981
14 7.46 2.51    3   38  1444     1  2 1982
15 6.81 2.77    3   39  1521     0  2 1983
16 5.44 1.43    2   40  1600     0  2 1984
17 5.08 1.72    1   42  1764     1  2 1985
18 5.85 1.86    1   42  1764     0  2 1986
19 7.69 1.83    1   43  1849     0  2 1987
20 7.63 1.79    0   44  1936     0  2 1988
```

Table 21.2. *Hours and Wages: Standard Linear Panel Model Estimators*[a]

	POLS	Between	Within	First Diff	RE–GLS	RE–MLE
α	7.442	7.483	7.220	.001	7.346	7.346
β	.083	.067	.168	.109	.119	.120
Robust se[b]	(.030)	(.024)	(.085)	(.084)	(.051)	(.052)
Boot se	[.030]	[.019]	[.084]	[.083]	[.056]	[.058]
Default se	{.009}	{.020}	{.019}	{.021}	{.014}	{.014}
R^2	.015	.021	.016	.008	.014	.014
RMSE	.283	.177	.233	.296	.233	.233
RSS	427.225	0.363	259.398	417.944	288.860	288.612
TSS	433.831	17.015	263.677	420.223	293.023	292.773
σ_α	.000		.181		.161	.162
σ_ε	.283		.232		.233	.233
λ	0.000	–	1.000	–	.585	.586
N	5320	532	5320	4788	5320	5320

[a] Shown are pooled OLS (POLS), between, within, first-differences, random effects (RE) GLS and MLE linear panel regression of lnhrs on lnwg. Standard errors for the slope coefficients are panel robust in parentheses, panel bootstrap in square brackets, and default estimates that assume iid errors in curly braces. The R^2, root mean square error (RMSE), residual sum of squares (RSS), total sum of squares (TSS), and sample size come from the appropriate regression given in Section 21.2. The parameter λ is defined after (21.11).

[b] se, standard error.

Their results

We will use lm to replicate the POLS, and plm to replicate POLS, within, and First Diff columns.

Pooled OLS

```
> mod1 = lm(lnhr ~ lnwg, data=mom)
> summary(mod1)

Call:
lm(formula = lnhr ~ lnwg, data = mom)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.441516   0.024126 308.438   <2e-16 ***
lnwg        0.082744   0.009125   9.068   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2834 on 5318 degrees of freedom
Multiple R-squared:  0.01523,	Adjusted R-squared:  0.01504
F-statistic: 82.22 on 1 and 5318 DF,  p-value: < 2.2e-16
```

	POLS	Between	Within	First Diff
α	7.442	7.483	7.220	.001
β	.083	.067	.168	.109
Robust se[b]	(.030)	(.024)	(.085)	(.084)
Boot se	[.030]	[.019]	[.084]	[.083]
Default se	{.009}	{.020}	{.019}	{.021}
R^2	.015	.021	.016	.008

- Here we find the same estimate for the coefficient and the default se.
- This standard error is likely wrong due to clustering. We will implement cluster-robust standard errors. Note these errors are also robust to heteroskedasticity.

- We can also run pooled OLS model directly in the plm package, and use their vcovHC function to correct the standard errors.

	POLS	Between	Within	First Diff
α	7.442	7.483	7.220	.001
β	.083	.067	.168	.109
Robust se[b]	(.030)	(.024)	(.085)	(.084)
Boot se	[.030]	[.019]	[.084]	[.083]
Default se	{.009}	{.020}	{.019}	{.021}
R^2	.015	.021	.016	.008

```
> #For PLM it is best to create a pdata.frame and indicate your
> #your year and grouping indices
> p.mom = pdata.frame(mom, index=c("id", "year"))
>
> mod1b = plm(lnhr ~ lnwg, data=p.mom, model="pooling")
> summary(mod1b)#mathches our OLS model
Oneway (individual) effect Pooling Model

Call:
plm(formula = lnhr ~ lnwg, data = p.mom, model = "pooling")

Balanced Panel: n=532, T=10, N=5320

Coefficients :
             Estimate Std. Error  t-value  Pr(>|t|)
(Intercept) 7.4415165  0.0241265 308.4379 < 2.2e-16 ***
lnwg        0.0827435  0.0091251   9.0677 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    433.83
Residual Sum of Squares: 427.23
R-Squared      :  0.015226
      Adj. R-Squared :  0.01522
F-statistic: 82.2223 on 1 and 5318 DF, p-value: < 2.22e-16

> #now if we want cluster-robust standard errors in PLM we
> #can use their following
> coeftest(mod1b,vcov=vcovHC(mod1b,cluster="group"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 7.441516   0.079505 93.5985 < 2.2e-16 ***
lnwg        0.082744   0.029241  2.8297  0.004676 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Within

- Here we ran a fixed effects model with the default standard errors and then used the vocHC function to produce the cluster-robust standard errors.

	POLS	Between	Within	First Diff
α	7.442	7.483	7.220	.001
β	.083	.067	.168	.109
Robust se[b]	(.030)	(.024)	(.085)	(.084)
Boot se	[.030]	[.019]	[.084]	[.083]
Default se	{.009}	{.020}	{.019}	{.021}
R^2	.015	.021	.016	.008

```
> mod3 = plm(lnhr ~ lnwg, data=p.mom, model="within")
> summary(mod3)
Oneway (individual) effect Within Model

Call:
plm(formula = lnhr ~ lnwg, data = p.mom, model = "within")

Balanced Panel: n=532, T=10, N=5320

Coefficients :
     Estimate Std. Error t-value  Pr(>|t|)
lnwg  0.16767    0.01887  8.8858 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    263.68
Residual Sum of Squares: 259.4
R-Squared      :  0.016227
      Adj. R-Squared :  0.014601
F-statistic: 78.9578 on 1 and 4787 DF, p-value: < 2.22e-16
> #now if we want cluster-robust standard errors in PLM we
> #can use their following
> coeftest(mod3,vcov=vcovHC(mod3,cluster="group"))

t test of coefficients:

     Estimate Std. Error t value Pr(>|t|)
lnwg 0.167675   0.084883  1.9754  0.04828 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Cameron and Trivedi use Stata to produce their results and Stata by default provides a constant term in their panel model output.
- The intercept that they report for the 'within' model is actually the average of the individual intercepts.
- You can see for yourself by calling:
  - mean(fixef(mod3)) = 7.219892

First Differenced

```
> mod4 = plm(lnhr ~ lnwg, data=p.mom, model="fd")
> summary(mod4)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = lnhr ~ lnwg, data = p.mom, model = "fd")

Balanced Panel: n=532, T=10, N=5320


Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
(intercept) 0.00082831 0.00427118  0.1939    0.8462
lnwg        0.10898515 0.02133514  5.1082 3.378e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    420.22
Residual Sum of Squares: 417.94
R-Squared      :  0.0054226
      Adj. R-Squared :  0.0054204
F-statistic: 26.0942 on 1 and 4786 DF, p-value: 3.378e-07

> #now if we want cluster-robust standard errors in PLM we
> #can use their following
> coeftest(mod4,vcov=vcovHC(mod4,cluster="group"))

t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(intercept) 0.00082831 0.00161311  0.5135   0.6076
lnwg        0.10898515 0.08363915  1.3030   0.1926
```

	POLS	Between	Within	First Diff
α	7.442	7.483	7.220	.001
β	.083	.067	.168	.109
Robust se[b]	(.030)	(.024)	(.085)	(.084)
Boot se	[.030]	[.019]	[.084]	[.083]
Default se	{.009}	{.020}	{.019}	{.021}
R^2	.015	.021	.016	.008

3. Panel Data and Dummy variable regression

Dummy variable regression

- The traditional view of a fixed effects model is to assume that the unobserved effect, a_i, is a parameter to be estimated for each *i*.
- Thus, in the following equation, a_i is the intercept for each observational unit.

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \quad t = 1,2,\dots.T$$

- The dummy variable regression resolves the fixed effects problem by including a dummy variable for each cross-sectional unit and then applying OLS to the model.
- Clearly, we cannot do this with cross-sectional data. There would be *N*+*k* parameters to estimate only *N* observations. Therefore, we need at least two time periods

- Estimating an intercept for each *i* is simple. Just add a dummy variable for each *i*.
- The dummy method gives us the exact same beta estimates as we would get from the time demeaned model and the standard error and other statistics are identical.
- By including the individual intercept terms, a_i, we have removed the variation that exists between the individual units and are left with only the within individual variation.

- The R-squared for this model is generally high because we are including a dummy for each cross-sectional unit which may explain much of the variation in the data.

- Note: We can use OLS on dummy variable regression and obtain correct standard errors as the df are properly adjusted due to the addition of a

Example of the NC Crime model with dummy variables

```
> lsdv = lm(lcrmrte ~ d82 + d83 + d84 + d85 + d86+ d87 + factor(county) + lpr
barr + lprbconv + lprbpris + lavgsen + lpolpc, data=nccrime)
> summary(lsdv)

Call:
lm(formula = lcrmrte ~ d82 + d83 + d84 + d85 + d86 + d87 + factor(county) +
    lprbarr + lprbconv + lprbpris + lavgsen + lpolpc, data = nccrime)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62833 -0.06629  0.00244  0.06955  0.53515

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.331667   0.169262  -7.867 2.05e-14 ***
d82                0.012580   0.021542   0.584 0.559471
d83               -0.079281   0.021340  -3.715 0.000225 ***
d84               -0.117728   0.021614  -5.447 7.87e-08 ***
d85               -0.111956   0.021846  -5.125 4.18e-07 ***
d86               -0.081827   0.021427  -3.819 0.000150 ***
d87               -0.040470   0.021039  -1.924 0.054945 .
factor(county)3   -0.498327   0.080759  -6.171 1.35e-09 ***
factor(county)5   -0.835923   0.076030 -10.995  < 2e-16 ***
factor(county)7   -0.261917   0.075348  -3.476 0.000551 ***
factor(county)9   -0.690049   0.079591  -8.670  < 2e-16 ***
factor(county)11  -0.979014   0.077611 -12.614  < 2e-16 ***
factor(county)13   0.021289   0.075005   0.284 0.776654
.
.
.
factor(county)193 -0.314364   0.075709  -4.152 3.84e-05 ***
factor(county)195 -0.062298   0.077478  -0.804 0.421715
factor(county)197 -0.696106   0.078340  -8.886  < 2e-16 ***
lprbarr           -0.359794   0.032419 -11.098  < 2e-16 ***
lprbconv          -0.285873   0.021217 -13.474  < 2e-16 ***
lprbpris          -0.182781   0.032461  -5.631 2.92e-08 ***
lavgsen           -0.004488   0.026447  -0.170 0.865315
lpolpc             0.424114   0.026366  16.086  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1387 on 529 degrees of freedom
Multiple R-squared:  0.9507,	Adjusted R-squared:  0.9414
F-statistic:   102 on 100 and 529 DF,  p-value: < 2.2e-16
```

Limits to Fixed Effect Models

- Fixed effect models cannot estimate coefficient on variables that do not vary for each unit.
  - De-meaned model shows why:

$$Y_{it} - \overline{Y}_{i\cdot} = \beta_1 (X_{it} - \overline{X}_{i\cdot}) + \tilde{\nu}_{it}$$

  - *Example*: in a panel data of individual opinion over time, a fixed effect model cannot estimate a coefficient on gender or race because for each individual in the panel, gender and race do not vary.

Variables that are fixed within unit

- We cannot estimate:

$$Crime_{it} = \beta_0 + \beta_1 Police_{it} + \beta_2 North_i + \alpha_i + \varepsilon_{it}$$

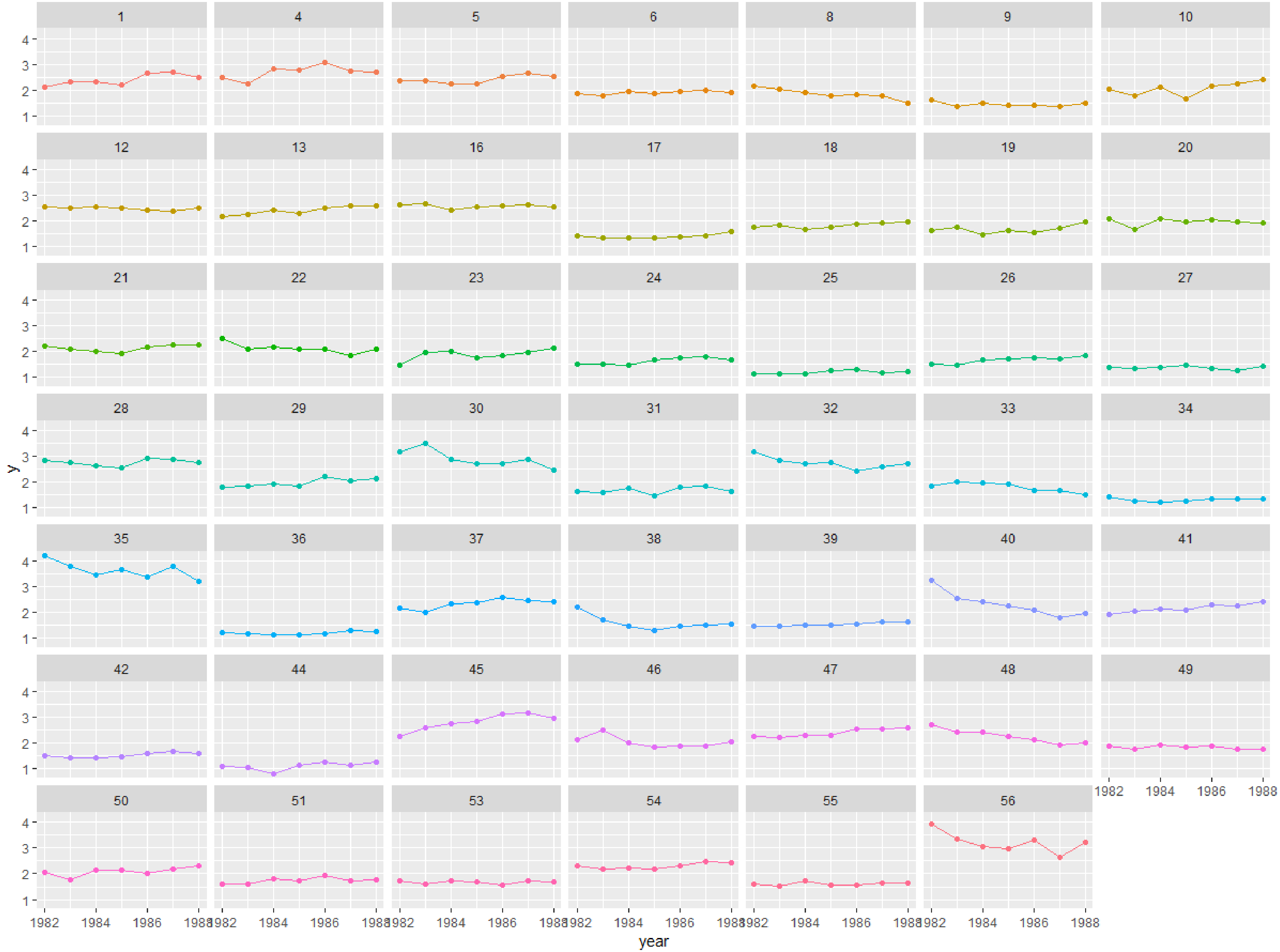
- We can, however, estimate

$$Crime_{it} = \beta_0 + \beta_1 Police_{it} + \beta_2 (Police_{it} \times North_i) + \alpha_i + \varepsilon_{it}$$

Let’s Walk through another example

Traffic fatalities and state beer taxes (Jackman 2010)

- Panel data set examining the link between traffic fatalities and beer taxes in the lower 48 states from 1982 to 1988.
- The dependent variable is the vehicle fatality rate (annually per 10,000 people).
- When we plot the data by state we can see that there is wide variation in the rates across states.



1
4
5
6
8
9
10
12
13
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
44
45
46
47
48
49
50
51
53
54
55
56
y
1
2
3
4
1982
1984
1986
1988
year

Let's begin with a naïve OLS estimate of the effect of beer tax on fatalities

$$fatality_{it} = b_0 + b_1 beertax_{it} + e_{it}$$

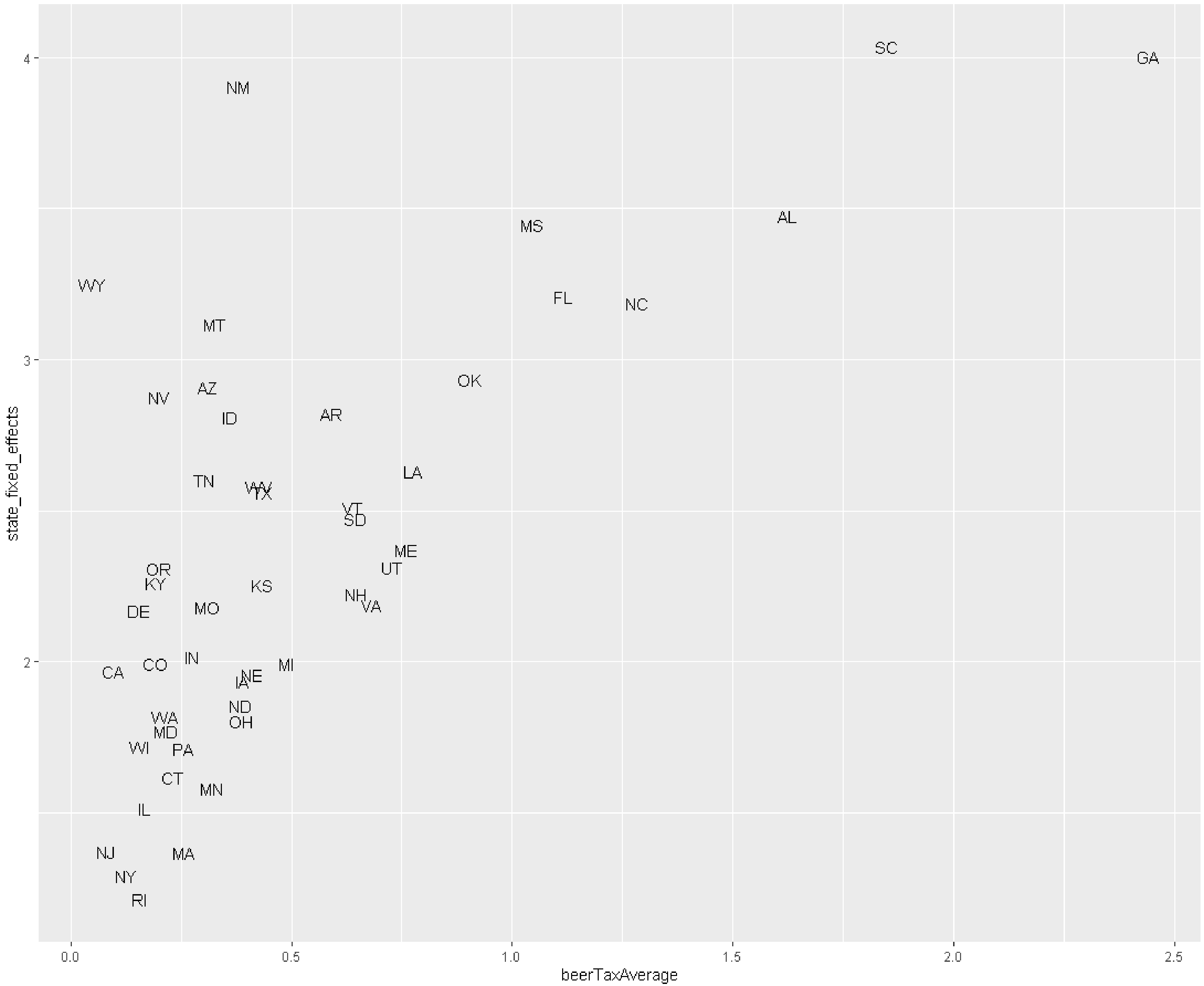
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.85331    0.04357  42.539  < 2e-16 ***
beertax      0.36461    0.06217   5.865 1.08e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- When we ignore the possible between-state heterogeneity, we get a surprising result.

Model using Fixed Effects

```
Coefficients :
        Estimate Std. Error t-value Pr(>|t|)
beertax -0.65587    0.18785 -3.4915 0.000556 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- When we properly deal with the unobservable fixed effects for each state we see a striking difference in the estimated effect of the beer tax.



state_fixed_effects
beerTaxAverage
4
3
2
0.0
0.5
1.0
1.5
2.0
2.5
SC
GA
NM
MS
AL
WY
FL
NC
MT
OK
NV
AZ
ID
AR
TN
WV
TX
LA
VT
SD
ME
UT
OR
KY
KS
NH
VA
DE
MO
CA
CO
IN
MI
NE
IA
ND
OH
WA
MD
WI
PA
CT
MN
IL
NJ
MA
NY
RI

Time for you to practice

- I have posted practice questions to this week's folder on Blackboard.
- The questions come directly from a prior methods comprehensive exam.
  - Also good practice for those doctoral students planning to take a methods exam.