Introduction to R Programming

Ozan Bakış[1]

[1] Bahcesehir University, Department of Economics and BETAM

Outline

1. Regression with cross-sectional data

2. Regression with cross-sections
   - Basics
   - DD estimator using Pooled CS

3. Regression with panel data
   - Basics
   - Estimating fixed effects model

Multiple regression: notation I

- linear regression model (estimated by OLS):

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i, \quad i = 1, \ldots, n.$$

- **Application:** estimation of wage equation using DCPS1988 data from AER (Applied Econometrics with R) package. CPS= Current Population Survey. ⇒ cross-section data on male workers (excluding self-employment and ed unpaid family work) aged 18 to 70 with positive annual income.

```r
f_url = "https://github.com/obakis/econ_data/raw/master/hls2011.rds"
download.file(url = f_url, destfile = "hls2011.rds", mode="wb")
hls = readRDS("hls2011.rds")
```

- Before regression let us look our variables of interest.

Multiple regression: notation II

```
head(hls,3)
```

```
##   id exper educ emp_sect emp_type hwage nuts1   wts urban female
## 1  1    33    2      pub   f-time  8.75    N9  45.8     1      0
## 2  2     2    2     priv   f-time  2.92   N12 178.8     0      1
## 3  3    22    5     priv   f-time  2.53    N2  66.9     1      1
```

```
vars = c("hwage","educ", "female","exper","emp_sect")
str(hls[,vars])
```

```
## 'data.frame': 762 obs. of  5 variables:
##  $ hwage   : num  8.75 2.92 2.53 58.33 3.89 ...
##  $ educ    : int  2 2 5 15 8 8 5 15 5 15 ...
##  $ female  : int  0 1 1 0 0 0 0 1 0 1 ...
##  $ exper   : int  33 2 22 21 16 49 22 6 17 2 ...
##  $ emp_sect: Factor w/ 3 levels "other","priv",..: 3 2 2 2 2 2 2 2 2 3 ...
```

Multiple regression: notation III

- Note that **emp_sect** is a **factor** variable with 12 levels. In R, categorical (nominal) and ordered categorical (ordinal) variables are called **factor**s. Each possible value of a categorical variable is called a level. In a regression a set of dummy variables will be automatically created by R. More precisely, if we have n groups/levels, $n-1$ dummy variables will be created.

```
r2e_1 = lm(log(hwage) ~ exper + I(exper^2) + educ + emp_sect, data=hls)
```

- Operators `+,-,:,*,/,^` have special meanings in a **formula** object. To ensure arithmetic meaning, we need either to protect by insulation in a function, e.g., `log(x1 * x2)` or to use `I()` function.

```
summary(r2e_1)
```

Multiple regression: notation IV

```
##
## Call:
## lm(formula = log(hwage) ~ exper + I(exper^2) + educ + emp_sect,
##     data = hls)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1542 -0.2863 -0.0271  0.2702  2.1929
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.51e-01   1.63e-01    1.54     0.12
## exper         3.17e-02   4.56e-03    6.96  7.6e-12 ***
## I(exper^2)   -4.93e-04   9.52e-05   -5.18  2.9e-07 ***
## educ          7.18e-02   4.57e-03   15.70  < 2e-16 ***
## emp_sectpriv  1.15e-01   1.52e-01    0.76     0.45
## emp_sectpub   7.59e-01   1.57e-01    4.84  1.5e-06 ***
```

Multiple regression: notation V

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.447 on 756 degrees of freedom
## Multiple R-squared:  0.569,Adjusted R-squared:  0.567
## F-statistic:  200 on 5 and 756 DF,  p-value: <2e-16
```

- Generic functions related to **`lm`** object (See **`help(lm)`** and **`names(cps.lm)`** for details):

`print()`	simple printed display
`summary()`	standard regression output
`coef()`	(or **`coefficients()`**) extract regression coefficients
`residuals()`	(or **`resid()`**) extract residuals
`fitted()`	(or **`fitted.values()`**) extract fitted values
`predict()`	predictions for new data
`plot()`	diagnostic plots
`confint()`	confidence intervals for the regression coefficients
`AIC()`	information criteria including AIC, BIC/SBC

Multiple regression: notation VI

- The `lm()` command, relies on `model.matrix()` for the creation of dummy variables.

```
dummy <- factor(LETTERS[1:4])
model.matrix( ~ dummy)
```

```
##   (Intercept) dummyB dummyC dummyD
## 1           1      0      0      0
## 2           1      1      0      0
## 3           1      0      1      0
## 4           1      0      0      1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$dummy
## [1] "contr.treatment"
```

- To change the base level of a factor variable (ex. ”region” variable) we can use `relevel` function

Multiple regression: notation VII

```
table(hls$emp_sect)
```

```
##
## other  priv   pub
##     9   557   196
```

```
levels(hls$emp_sect)
```

```
## [1] "other" "priv"  "pub"
```

```
contrasts(hls$emp_sect) #other is base level
```

```
##       priv pub
## other    0   0
## priv     1   0
## pub      0   1
```

```
hls$emp_sect <- relevel(hls$emp_sect, ref = "pub")
r2e_2 <- update(r2e_1, formula = . ~ .) ## we change nothing here!
summary(r2e_2)$coef
```

Multiple regression: notation VIII

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.009864   9.47e-02   10.66 8.14e-25
## exper           0.031745   4.56e-03    6.96 7.62e-12
## I(exper^2)     -0.000493   9.52e-05   -5.18 2.86e-07
## educ            0.071799   4.57e-03   15.70 2.70e-48
## emp_sectother  -0.758748   1.57e-01   -4.84 1.54e-06
## emp_sectpriv   -0.643402   4.36e-02  -14.74 1.98e-43
```

`update()` is used for updating an `lm` object. Since we do not change the LHS or the RHS of the `formula`, above our goal is just re-doing the same regression with new base level for `region` variable.

- What if we want to add or remove some variables

```
r2e_3 <- update(r2e_2, formula = . ~ . - emp_sect)
summary(r2e_3)$coef
```

Multiple regression: notation IX

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.014760   0.075683   0.195 8.45e-01
## exper        0.046859   0.005032   9.313 1.32e-19
## I(exper^2)  -0.000696   0.000106  -6.550 1.06e-10
## educ         0.107217   0.004419  24.263 1.11e-96
```

```
r2e_4 <- update(r2e_3, formula = . ~ . + female)
summary(r2e_4)$coef
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.014430   0.076451  0.1887 8.50e-01
## exper        0.046877   0.005068  9.2504 2.24e-19
## I(exper^2)  -0.000696   0.000107 -6.5258 1.24e-10
## educ         0.107196   0.004472 23.9696 6.51e-95
## female       0.001440   0.045617  0.0316 9.75e-01
```

```
confint(r2e_4) # by default: level = 0.95
```

Multiple regression: notation X

```
##                  2.5 %    97.5 %
## (Intercept) -0.135650  0.164510
## exper        0.036929  0.056825
## I(exper^2)  -0.000905 -0.000486
## educ         0.098417  0.115975
## female      -0.088111  0.090990
```

```
confint(r2e_4, level=0.9)
```

```
##                    5 %     95 %
## (Intercept) -0.111474  0.14033
## exper        0.038531  0.05522
## I(exper^2)  -0.000871 -0.00052
## educ         0.099831  0.11456
## female      -0.073685  0.07656
```

Interactions I

Formula	Description
`y~a+x`	Model without interaction: identical slopes with respect to `x` but different intercepts with respect to `a`.
`y~a*x` `y~a+x+a:x`	Model with interaction: the term `a:x` gives the difference in slopes compared with the reference category.
`y~a/x` `y~a+x%in%a`	Model with interaction: produces the same fitted values as the model above but using a nested coefficient coding. An explicit slope estimate is computed for each category in `a`.

Interactions II

```r
#install.packages("lmtest")
library(lmtest) # for inference
# need to convert female into factor variable
hls$female=factor(hls$female)

## main effects + interaction
r2e_5 = lm(log(hwage) ~ exper + I(exper^2) + educ*female, data=hls)
coeftest(r2e_5)
## nested models
r2e_6 = lm(log(hwage) ~ female/(0+exper + I(exper^2) + educ), data=hls)
coeftest(r2e_6)
```

F test: linear restrictions I

- Consider the following model

```
coeftest(r2e_4)
```

```
## Error in coeftest(r2e_4): could not find function "coeftest"
```

- We want to test $H_0 : \beta_3 = 0.07, \beta_4 = 0$. These are called exclusion restrictions.

```
#install.packages("car")
library(car)
```

```
## Loading required package:  carData
```

```
linearHypothesis(r2e_4, c("educ=0.07","female=0")) # reject null
```

F test: linear restrictions II

```
## Linear hypothesis test
##
## Hypothesis:
## educ = 0.07
## female = 0
##
## Model 1: restricted model
## Model 2: log(hwage) ~ exper + I(exper^2) + educ + female
##
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1    759 213
## 2    757 194  2      18.2 35.4  2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(r2e_4, "educ - 2*exper  = 0") # cannot reject null
```

F test: linear restrictions III

```
## Linear hypothesis test
##
## Hypothesis:
## - 2 exper  + educ = 0
##
## Model 1: restricted model
## Model 2: log(hwage) ~ exper + I(exper^2) + educ + female
##
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1    758 195
## 2    757 194  1     0.441 1.72   0.19
```

Heteroskedasticity robust std. erros I

```r
library(lmtest) # for coeftest
coeftest(r2e_4) # assuming homoskedasticity
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.014430   0.076451    0.19     0.85
## exper        0.046877   0.005068    9.25  < 2e-16 ***
## I(exper^2)  -0.000696   0.000107   -6.53  1.2e-10 ***
## educ         0.107196   0.004472   23.97  < 2e-16 ***
## female       0.001440   0.045617    0.03     0.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heteroskedasticity robust std. erros II

```
library(sandwich) # for vcovHC
coeftest(r2e_4, vcov = vcovHC) # heteroskedasticity robust, R default: "HC3"
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.014430   0.079666    0.18     0.86
## exper        0.046877   0.005106    9.18  < 2e-16 ***
## I(exper^2)  -0.000696   0.000108   -6.42  2.4e-10 ***
## educ         0.107196   0.005053   21.21  < 2e-16 ***
## female       0.001440   0.044464    0.03     0.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heteroskedasticity robust std. erros III

```
coeftest(r2e_4, vcov = vcovHC(r2e_4, "HC3")) # robust, R default: "HC3"
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.014430   0.079666    0.18     0.86
## exper        0.046877   0.005106    9.18  < 2e-16 ***
## I(exper^2)  -0.000696   0.000108   -6.42  2.4e-10 ***
## educ         0.107196   0.005053   21.21  < 2e-16 ***
## female       0.001440   0.044464    0.03     0.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heteroskedasticity robust std. erros IV

```
coeftest(r2e_4, vcov = vcovHC(r2e_4, "HC1")) # robust, Stata default: "HC1"
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.014430   0.078957    0.18     0.86
## exper        0.046877   0.005002    9.37  < 2e-16 ***
## I(exper^2)  -0.000696   0.000105   -6.62  6.7e-11 ***
## educ         0.107196   0.005022   21.34  < 2e-16 ***
## female       0.001440   0.044203    0.03     0.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Outline

1. Regression with cross-sectional data

2. Regression with cross-sections
   - Basics
   - DD estimator using Pooled CS

3. Regression with panel data
   - Basics
   - Estimating fixed effects model

Pooled cross section: overview I

- Data obtained by pooling cross sections (PCS) are very useful for establishing trends and conducting policy analysis.
- A pooled cross section is available whenever a survey is repeated over time with new random samples obtained in each time period.
- Examples include the Current Population Survey (CPS) in USA and Household Labor Survey (Hanehalkı İşgücü Anketi) in Turkey.
- With a PCS, often a goal is to see how the mean value of a variable (fertility) has changed over time in ways that cannot be explained by observable variables (education).
- **Ex.**: Has the fertility rate changed in ways that cannot be explained by education?

Pooled cross section: overview II

- From a policy perspective, PCSs are at the foundation of *difference-in-differences* estimation.
- The typical DD setup is that data can be collected both before and after an intervention (or ”treatment”), and there is (at least) one ”control group” and (at least) one ”treatment” group.
- Often the intervention is of a yes/no form. But other nonbinary treatments (such as class size) can be handled, too.

Application I

- IS the Change in women's fertility in the USA (1972-1984) can be explained by rise in education levels of women?
- How much of the fall in average fertility cannot be explained by changes in observed factors, including education? Here we require a PCS and look at coefficients on year dummies.
- How much of the overall fall in average fertility be explained by increases in average education?
- Before a full regression model let us go step by step and try to understand the underlying patterns. What is the trend for average number of kids over years?

Application II

```r
f_url = "https://github.com/obakis/econ_data/raw/master/fertil1.rds"
download.file(url = f_url, destfile = "fertil1.rds", mode="wb")
dat = readRDS("fertil1.rds")

#with(dat, tapply(kids, year, FUN=summary))
aggregate(kids ~ year, FUN=mean, data=dat)
```

```
##   year kids
## 1   72 3.03
## 2   74 3.21
## 3   76 2.80
## 4   78 2.80
## 5   80 2.82
## 6   82 2.40
## 7   84 2.24
```

- The average fertility rate fell by about 0.79, $-0.79 = 2.24 - 3.03$.

Application III

- The same could be done through a regression as well.

```
reg1 = lm(kids ~year, data=dat) # this is probably not what we want
coeftest(reg1) # or maybe it is???
```

```
## 
## t test of coefficients:
## 
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   8.4256     0.9267    9.09  < 2e-16 ***
## year         -0.0727     0.0118   -6.14  1.1e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
dat$year_f = factor(dat$year)
reg2 = lm(kids ~ year_f, data=dat) # this is probably a better way
coeftest(reg2) # You see a trend over time?
```

Application IV

```
## 
## t test of coefficients:
## 
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    3.026      0.130   23.24  < 2e-16 ***
## year_f74       0.182      0.180    1.02  0.30976    
## year_f76      -0.223      0.185   -1.20  0.22912    
## year_f78      -0.221      0.188   -1.18  0.23975    
## year_f80      -0.209      0.189   -1.11  0.26865    
## year_f82      -0.622      0.177   -3.53  0.00044 ***
## year_f84      -0.788      0.179   -4.41  1.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Application V

- Again we can say that the average fertility rate fell by about 0.79. But more importantly we have an idea about its significance. This is the benefit of regression compared to simple comparison of means!
- Education is an important determinant of fertility. Let us see how it changes over years

```
aggregate(educ ~ year, FUN=mean, data=dat)
```

```
##   year educ
## 1   72 12.2
## 2   74 12.3
## 3   76 12.2
## 4   78 12.6
## 5   80 12.9
## 6   82 13.2
## 7   84 13.3
```

Application VI

```
reg3 = lm(educ ~ year_f, data=dat) # this is probably OK
coef(summary(reg3)) # You see a trend over time?
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.1538      0.209  58.164 0.000000
## year_f74      0.1467      0.288   0.509 0.610709
## year_f76      0.0764      0.297   0.257 0.797296
## year_f78      0.4895      0.302   1.620 0.105495
## year_f80      0.7264      0.303   2.400 0.016566
## year_f82      1.0720      0.283   3.783 0.000163
## year_f84      1.1117      0.287   3.879 0.000111
```

- Overall the increase in mean eduaction is 1.11 years. To see the effect of this increase on fertility we need to know the partial effect of education on fertility. For this we run the following regression

```
dat$year_f = factor(dat$year)
reg4 <- lm(kids~educ+age+agesq+black+east+northcen+west+year_f, data=dat)
coef(summary(reg4))
```

Application VII

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.95228    3.05004  -2.607 9.25e-03
## educ        -0.12269    0.01803  -6.806 1.64e-11
## age          0.53904    0.13837   3.896 1.04e-04
## agesq       -0.00588    0.00156  -3.762 1.78e-04
## black        1.09095    0.17311   6.302 4.22e-10
## east         0.25290    0.12685   1.994 4.64e-02
## northcen     0.38523    0.11853   3.250 1.19e-03
## west         0.23257    0.16532   1.407 1.60e-01
## year_f74     0.25608    0.17265   1.483 1.38e-01
## year_f76    -0.10630    0.17857  -0.595 5.52e-01
## year_f78    -0.07050    0.18136  -0.389 6.98e-01
## year_f80    -0.07855    0.18261  -0.430 6.67e-01
## year_f82    -0.53255    0.17230  -3.091 2.05e-03
## year_f84    -0.54226    0.17436  -3.110 1.92e-03
```

Application VIII

- Each additional year of education is estimated to reduce the number of children by about 0.123, on average.
- Compared to 1972, fertility fell by about 0.55 children in 1984. This is the drop that cannot be explained by the explanatory variables.
- Of the overall drop of about 0.79 children, the increase in education (1.11 years on average) accounts for about $0.14 \approx 1.11 \times 0.123$ of that, or about 18%.

- In the previous estimation with the fertility data, we assumed the effect of education (and all other variables) was the same over time.
- We can easily allow the slopes to change over time by forming interactions and adding them to the model.

Application IX

```
reg5 = lm(kids~educ+age+agesq+black+east+northcen+west+
            year_f+year_f:educ, data=dat)
coef(summary(reg5))
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -8.61295    3.12654  -2.755 5.97e-03
## educ           -0.02456    0.05357  -0.458 6.47e-01
## age             0.51493    0.13894   3.706 2.21e-04
## agesq          -0.00561    0.00157  -3.573 3.67e-04
## black           1.09031    0.17333   6.290 4.55e-10
## east            0.24805    0.12711   1.952 5.12e-02
## northcen        0.37446    0.11869   3.155 1.65e-03
## west            0.21657    0.16582   1.306 1.92e-01
## year_f74        0.82860    0.90059   0.920 3.58e-01
## year_f76        0.89183    0.87799   1.016 3.10e-01
## year_f78        1.71434    0.94992   1.805 7.14e-02
## year_f80        0.97952    0.89447   1.095 2.74e-01
```

Application X

```
## year_f82        1.06195    0.87204   1.218 2.24e-01
## year_f84        1.54243    0.89484   1.724 8.50e-02
## educ:year_f74 -0.04779    0.07230  -0.661 5.09e-01
## educ:year_f76 -0.08241    0.07058  -1.168 2.43e-01
## educ:year_f78 -0.14509    0.07515  -1.931 5.38e-02
## educ:year_f80 -0.08768    0.07024  -1.248 2.12e-01
## educ:year_f82 -0.12869    0.06807  -1.891 5.89e-02
## educ:year_f84 -0.16553    0.06961  -2.378 1.76e-02
```

Let us test year and education interaction terms:

```
library(car)
linearHypothesis(reg5, matchCoefs(reg5, "educ:year"), vcov=hccm(reg5))
```

Application XI

```
## Linear hypothesis test
##
## Hypothesis:
## educ:year_f74 = 0
## educ:year_f76 = 0
## educ:year_f78 = 0
## educ:year_f80 = 0
## educ:year_f82 = 0
## educ:year_f84 = 0
##
## Model 1: restricted model
## Model 2: kids ~ educ + age + agesq + black + east + northcen + west +
##     year_f + year_f:educ
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df    F Pr(>F)
```

Application XII

```
## 1   1115
## 2   1109  6 1.02   0.41
```

Jointly insignificant, even though `educ:year_f84` and `educ:year_f82` are individually significant.

- Coefficient on, say, `year_f84` is the difference in fertility between 1984 and 1972 at $educ = 0$; not interesting.
- Effect of schooling in base year very close to zero.
- The joint test for all interactions with $educ$ gives p-value = 0.41, so we cannot reject the null that the effect of education has been constant. But it seems fertility has become more sensitive to education in the last couple of years of the data (1982, 1984).

DD with 2 groups and 2 time periods I

- Useful to study the data coming from a **natural experiment** (or a quasi-experiement). This is called natural experiment because
  ⇒ an exogenous event (usually a change in government policy) changes the conditions under which individuals / firms etc. operate.
  ⇒ There are at least one control group (not affected by policy change) and one treatment group (affected by policy change).

- DD methodology is used widely to evaluate the consequences of natural experiments (or quasi-experiments). There are two key elements in natural experiments:

DD with 2 groups and 2 time periods II

- Outcomes are observed for two groups over two time periods. One of the groups is exposed to a ”treatment” (or intervention) in the second period but not in the first period. The second group is not exposed to the treatment during either period.

- In the textbook version where we have 2 groups and 2 time periods. As a result, treating one of groups and time periods as reference, we need at least two dummy variables: 1 for the other group and 1 for the other time period.

- Let A be the control group and B the treatment group. Let $dT = 1$ for a unit in B, and $dT = 0$ for a unit in A in both periods. Similarly, let $d2 = 1$ for a unit in the second time period and $d2 = 0$ for a unit in the first period for both states.

DD with 2 groups and 2 time periods III

- **Ex.:** Remember the famous minimum wage example: Card and Krueger (1994) studied the increase in the minimum wage in New Jersey from 4.25 USD to 5.05 USD. This change took effect on April 1, 1992. The minimum wage in Pennsylvania remained at 4.25 USD throughout this period.
- Card and Krueger collected data on employment at fast food restaurants in New Jersey and Pennsylvania (the neighboring state) in February (before) and in November (after).
- Here New Jersey is the ”treatment state” (B) and Pennsylvania is the ”control state” (A). Without data from Pensylvania, we cannot control for aggregate changes over time that affect employment in both states.

DD with 2 groups and 2 time periods IV

A typical model would be:

$$y = \beta_0 + \beta_1 dT + \delta_0 d2 + \delta_1 d2 \cdot dT + u \tag{1}$$

where y is the outcome of interest. What was missing in previous regression is the specifity of the second period for treatment group !

	Before (1)	After (2)	After $-$ Before
Control (A)	β_0	$\beta_0 + \delta_0$	δ_0
Treatment (B)	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment $-$ Control	β_1	$\beta_1 + \delta_1$	δ_1

We could, in principle, use the averages directly :

DD with 2 groups and 2 time periods V

$$\begin{aligned}\hat{\delta}_{DD} = \hat{\delta}_1 &= (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1}) \\ &= (\bar{y}_{B,2} - \bar{y}_{A,2}) - (\bar{y}_{B,1} - \bar{y}_{A,1})\end{aligned}$$

Explain why the following models are not useful for a DD analysis?

- Use only ”after” data and estimate

$$y = \beta_0 + \beta_1 dT + u \tag{2}$$

DD with 2 groups and 2 time periods VI

- β_1 can be interpreted as the effect of minimum wage increase. BUT: Remember that β_1 has a *ceteris paribus* / causal interpretation only when it is possible to keep other factors fixed so that $\Delta u = 0$. For this we need the **zero conditional mean** assumption

  $$\mathsf{Cor}(dT, u) = 0 \Rightarrow E(u|dT) = 0$$

  All unobserved factors should be similar for NJ and PA, which is not likely.

- Use only New Jersey data and estimate

  $$y = \beta_0 + \beta_1 d2 + u \tag{3}$$

DD with 2 groups and 2 time periods VII

- β_1 can be interpreted as the effect of minimum wage increase. BUT: the **zero conditional mean** assumption in this case:

  $$\text{Cor}(d2, u) = 0 \Rightarrow E(u|d2) = 0$$

  For above to be a good estimate we need that nothing changes over time except the treatment / minimum wage increase. Any other change should not affect the employment level in NJ, again, which is not likely. In real life, many things are changing over time.

- Use only both states and both periods and estimate

  $$y = \beta_0 + \beta_1 dT + \delta_0 d2 + u \tag{4}$$

DD with 2 groups and 2 time periods VIII

- This regression allows us to know wheter B is different from A (in which period?) and wheter $d2$ is different from $d1$ (for which state?). What is missing?
- What is missing is ”common (or parallel) trends assumption”. We may assume that anything that is changing over time is common to both (control and treatment) states. Then only we can decompose total effect into ”treatment effect” and ”all other effects”.
- We need to decompose the treatment and time effects!

$$y = \beta_0 + \beta_1 dT + \delta_0 d2 + \delta_1 d2 \cdot dT + u \tag{5}$$

- While powerful, the basic DD approach can suffer from several problems.

DD with 2 groups and 2 time periods IX

- First, there may be compositional effects. For example, when studying the effects of the intervention to reduce class size, it may be the case that the students in the two years are not comparable. Perhaps the smaller class sizes attracts new students to that district. This is the problem of *compositional changes*.
  $\Rightarrow$ Can control for changes in composition to some extent by including observed regressors as controls.

$$y = \beta_0 + \beta_1 dT + \delta_0 d2 + \delta_1 d2 \cdot dT + \mathbf{x}\gamma + u$$

DD with 2 groups and 2 time periods X

- Second, a potential problem with using only two periods is that the control and treatment groups may be trending at different rates having nothing to do with the intervention. There is a large literature on dealing with violation of parallel trends assumption.
  ⇒ Only way to solve this problem is get another control group or more years of data.

Application I

Effect of a Garbage Incinerator's Location on Housing Prices

- the effect that a new garbage incinerator had on housing values in North Andover, Massachusetts. The rumor that a new incinerator would be built in North Andover began after 1978, and construction began in 1981.
- The incinerator was expected to be in operation soon after the start of construction. We will use data on prices of houses that sold in 1978 and another sample on those that sold in 1981.
- Year dummy is $y81$ it takes the value of 1 for year 1981 and 0 for ear 1978. We also define $nearinc = 1$ if a house is near the incinerator, more concertely if it is within three miles.

Application II

```r
f_url = "https://github.com/obakis/econ_data/raw/master/kielmc.rds"
download.file(url = f_url, destfile = "kielmc.rds", mode="wb")
dat = readRDS("kielmc.rds")

xtabs(~year,dat)
```

```
## year
## 1978 1981
##  179  142
```

```r
aggregate(rprice~year, FUN=mean, data=dat)
```

```
##   year rprice
## 1 1978  76628
## 2 1981  92663
```

Application III

- We measure all housing prices in 1978 dollars, using the Boston housing price index. Let *rprice* denote the house price in real terms. A naive analyst would use only the 1981 data and estimate a very simple model:

$$rprice = \delta_0 + \delta_1 nearinc + u$$

```
reg1 = lm(rprice ~ nearinc, data=subset(dat, year==1981))
coef(summary(reg1))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   101308       3093   32.75 1.65e-67
## nearinc       -30688       5828   -5.27 5.14e-07
```

Application IV

- Since this is a simple regression on a single dummy variable, the intercept is the average selling price for homes not near the incinerator, and the coefficient on nearinc is the difference in the average selling price between homes near the incinerator and those that are not. The estimate shows that the average selling price for the former group was 30688 USD less than for the latter group.
- Unfortunately, the SRM above does not imply that the siting of the incinerator is causing the lower housing values. In fact, if we run the same regression for 1978 (before the incinerator was even rumored), we obtain

```r
reg2 = lm(rprice ~ nearinc, data=subset(dat, year==1978))
coef(summary(reg2))
```

Application V

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82517       2654   31.09 1.24e-73
## nearinc       -18824       4745   -3.97 1.05e-04
```

- Therefore, even before there was any talk of an incinerator, the average value of a home near the site was 18824 USD less than the average value of a home not near the site. This is consistent with the view that the incinerator was built in an area with lower housing values.

- If so, we would expect a negative relationship found in the simple regression even if the new incinerator had no effect on housing prices.

Application VI

- How, then, can we tell whether building a new incinerator depresses housing values? The key is to look at how the coefficient on *nearinc* changed between 1978 and 1981. The difference in average housing value was much larger in 1981 than in 1978 (30688 vs 18824), even as a percentage of the average value of homes not near the incinerator site. The difference in the two coefficients on nearinc is

$$\hat{\delta}_1 = 30688 - 18824 = 11864$$

Application VII

- This is our estimate of the effect of the incinerator on values of homes near the incinerator site. In empirical economics, $\hat{\delta}_1$ has become known as the difference-in-differences estimator because it can be expressed as

$$\hat{\delta}_1 = (\bar{rprice}_{81,nr} - \bar{rprice}_{81,fr}) - (\bar{rprice}_{78,nr} - \bar{rprice}_{78,fr})$$

  where fr means ”farther away from the incinerator site” and nr means ”near the incinerator site”.

- An important question is whether this estimate is different from zero? For this, we need std. error of it. How can we compute a std. error for $\hat{\delta}_1$?

Application VIII

- Actually this can be obtained through the following model

$$rprice = \beta_0 + \delta_0\, y81 + \beta_1\, nearinc + \delta_1\, nearinc \cdot y81 + u$$

Let us look at the meaning of all coefficients:

  - The intercept, β_0, is the average price of a home not near the incinerator in 1978.
  - The parameter δ_0 captures changes in all housing values in North Andover from 1978 to 1981.
  - The coefficient on nearinc, β_1, measures the location effect that is not due to the presence of the incinerator: even in 1978, homes near the incinerator site sold for less than homes farther away from the site.

Application IX

- The parameter of interest is, δ_1, the coefficient on the interaction term $y81 \cdot nearinc$. This measures the decline in housing values due to the new incinerator, provided we assume that houses both near and far from the site did not appreciate at different rates for other reasons.

```r
#did <- lm(rprice ~ y81 + nearinc + nearinc:y81, data=kielmc)
did <- lm(rprice ~ nearinc*y81, data=dat)
coef(summary(did))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82517       2727   30.26 1.71e-95
## nearinc       -18824       4875   -3.86 1.37e-04
## y81            18790       4050    4.64 5.12e-06
## nearinc:y81   -11864       7457   -1.59 1.13e-01
```

Application X

- If desired, one can add other control variables in the same way as we would for regular multiple regression model. The benefit in adding more control variables is that this may control for the change in average house attributes (quality, space etc.). Even if this does not change that much, including house characteristics can greatly reduce the error variance, which can then shrink the standard error of the parameter of interest.

```
#intst: dist. to interstate, ft
#bath: number of bathrooms
#rooms: number of rooms
did_ctr <- lm(log(rprice)~nearinc*y81+age+I(age^2)+log(intst)+
                            log(land)+log(area)+rooms+baths, data=dat)
coef(summary(did_ctr))
```

Application XI

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.65e+00   4.16e-01  18.399 1.29e-51
## nearinc      3.22e-02   4.75e-02   0.679 4.98e-01
## y81          1.62e-01   2.85e-02   5.687 2.99e-08
## age         -8.36e-03   1.41e-03  -5.924 8.37e-09
## I(age^2)     3.76e-05   8.67e-06   4.342 1.92e-05
## log(intst)  -6.14e-02   3.15e-02  -1.950 5.21e-02
## log(land)    9.98e-02   2.45e-02   4.077 5.81e-05
## log(area)    3.51e-01   5.15e-02   6.813 4.98e-11
## rooms        4.73e-02   1.73e-02   2.732 6.66e-03
## baths        9.43e-02   2.77e-02   3.400 7.61e-04
## nearinc:y81 -1.32e-01   5.20e-02  -2.531 1.19e-02
```

Comparing both:

Application XII

	Dependent variable:	
	rprice	log(rprice)
	(1)	(2)
nearinc	−18,824.000*** (4,875.000)	0.032 (0.047)
y81	18,790.000*** (4,050.000)	0.162*** (0.028)
age		−0.008*** (0.001)
I(age^2)		0.00004*** (0.00001)
log(intst)		−0.061* (0.032)
log(land)		0.100*** (0.024)
log(area)		0.351*** (0.051)
rooms		0.047*** (0.017)
baths		0.094*** (0.028)
nearinc:y81	−11,864.000 (7,457.000)	−0.132** (0.052)
Constant	82,517.000*** (2,727.000)	7.650*** (0.416)
Observations	321	321
Adjusted R^2	0.166	0.724
Note:		*p$<$0.1; **p$<$0.05; ***p$<$0.01

Application XIII

Outline

1. Regression with cross-sectional data

2. Regression with cross-sections
   - Basics
   - DD estimator using Pooled CS

3. Regression with panel data
   - Basics
   - Estimating fixed effects model

Storing panel data I

- The best way to store panel data is to stack the time periods for each i on top of each other. In particular, the time periods for each unit should be adjacent, and stored in chronological order (from earliest period to the most recent). This is sometimes called the ”long” storage format. It is by far the most common.

```
f_url = "https://github.com/obakis/econ_data/raw/master/wagepan.rds"
download.file(url = f_url, destfile = "wagepan.rds", mode="wb")
dat = readRDS("wagepan.rds")
```

- While not absolutely necessary for some procedures, it is best to tell R that you have a panel data set. We can convert a regular data frame into a panel data frame using

```
pdat = pdata.frame(dat, index=c("i_var","t_var"))
```

  In our example

Storing panel data II

```r
head(dat,3)
```

```
##   nr year agric black bus construc ent exper fin hisp poorhlth hours
## 1 13 1980     0     0   1        0   0     1   0    0        0  2672
## 2 13 1981     0     0   0        0   0     2   0    0        0  2320
## 3 13 1982     0     0   1        0   0     3   0    0        0  2940
##   manuf married min nrthcen nrtheast occ1 occ2 occ3 occ4 occ5 occ6
## 1     0       0   0       0        1    0    0    0    0    0    0
## 2     0       0   0       0        1    0    0    0    0    0    0
## 3     0       0   0       0        1    0    0    0    0    0    0
##   occ7 occ8 occ9 per pro pub rur south educ tra trad union lwage d81
## 1    0    0    1   0   0   0   0     0   14   0    0     0  1.20   0
## 2    0    0    1   1   0   0   0     0   14   0    0     1  1.85   1
## 3    0    0    1   0   0   0   0     0   14   0    0     0  1.34   0
##   d82 d83 d84 d85 d86 d87 expersq
## 1   0   0   0   0   0   0       1
## 2   0   0   0   0   0   0       4
## 3   1   0   0   0   0   0       9
```

Storing panel data III

```
vars = c("nr", "year", "exper", "hours", "educ", "lwage")
#install.packages("plm")
library(plm)
pdat = pdata.frame(dat, index=c("nr","year"))
head(pdat[,vars])
```

```
##         nr year exper hours educ lwage
## 13-1980 13 1980     1  2672   14  1.20
## 13-1981 13 1981     2  2320   14  1.85
## 13-1982 13 1982     3  2940   14  1.34
## 13-1983 13 1983     4  2960   14  1.43
## 13-1984 13 1984     5  3071   14  1.57
## 13-1985 13 1985     6  2864   14  1.70
```

```
pdim(pdat)
```

```
## Balanced Panel: n = 545, T = 8, N = 4360
```

Estimating fixed effects model: FD I

- If the explanatory variable changes over time – at least for some units in the population – heterogeneity bias can be solved by differencing away a_i in the model. For this, write the time periods in reverse order for any unit i:

$$\begin{aligned} y_{i2} &= (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \\ y_{i1} &= \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \end{aligned}$$

  Subtract time period one from time period two to get

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

Estimating fixed effects model: FD II

- If we define $\Delta y_i = y_{i2} - y_{i1}$, where $\Delta = \textit{change}$ (or *difference*) (similarly for Δx_i and Δu_i) we get the following estimating equation

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

  from which we derive the *first-difference estimator*. (With more than two time periods, other orders of differencing are possible; hence the qualifier "first" .) We will refer to the *FD estimator*.

- Differencing away the unobserved effect, a_i, is simple but can be very powerful for isolating causal effects. In estimating equation we need, among other Gauss-Markov assumptions, that Δu being uncorrelated with Δx.

Estimating fixed effects model: FD III

- Important: β_1 is the original coefficient we are interested in. We compute it using the estimating equation

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

  but interpret it as if we have estimated it in levels equation (FE model)

  $$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}$$

  and controlled for a_i.

- Notice that the intercept in the differenced equation, is the *change* in the intercept over the two time periods. It is sometimes interesting to study this change.

- If $\Delta x_i = 0$ for all i, or even if Δx_i is the same nonzero constant, this strategy does not work. We need some variation in Δx_i across i.

Estimating fixed effects model: FD IV

- **Example:** Effects of City Unemployment Rates on Crime Rates
- Once we create a panel data set, it is easy to create variables for changes, growth rates, lagged values etc. Below, we will create two variables for ”change in crime rate” and ”change in unemployment rate” between years 1982 and 1987.

```
f_url = "https://github.com/obakis/econ_data/raw/master/crime2.rds"
download.file(url = f_url, destfile = "crime2.rds", mode="wb")
dat = readRDS("crime2.rds")

library(plm)
pdat = pdata.frame(dat, index=46) # n= number of cities each year
### FD estimator using basic lm: we need to define first differences manually
pdat$d_crmrte = diff(pdat$crmrte) # pdat$crmrte - lag(pdat$crmrte)
pdat$d_unem = diff(pdat$unem) # pdat$unem - lag(pdat$unem)
reg2_fd1 = lm(d_crmrte ~ d_unem, data=pdat)
coef(summary(reg2_fd1))
```

Estimating fixed effects model: FD V

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.40      4.702    3.28  0.00206
## d_unem          2.22      0.878    2.53  0.01519
```

```
## FD estimator using plm package. Note: we use same covariates as pooling
# reg2_fd2 = plm(crmrte ~ d87 + unem, data=pdat, model = "fd")
# coef(summary(reg2_fd2))
```

- Note that *d87* is dropped in the FD regression and the intercept is interpreted as the change in the intercept over the two time periods.
- The intercept is equal to 15.40 which means that, if the unemployment rate did not change, the crime rate would be predicted to increase by about 15 crimes per 1,000 people.

Estimating fixed effects model: FD VI

- The coefficient on `d_unem` is statistically significant and of the expected sign: a one percentage point increase in the unemployment rate increases the crime rate by about 2.2 crimes per 1000 people.

Estimating fixed effects model: FE I

- When we believe that $Cov(x_{it}, a_i) \neq 0$, differencing is only one method of eliminating a_i (which itself is sometimes called a *fixed effect*).
- Alternatively, can use the ”fixed effects” or ”within” transformation: remove the within i time averages.
- Consider the simple model

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$$

- Average this equation across t to get

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + a_i + \bar{u}_i$$

  where $\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$ is a ”time average” for unit i. Similarly for $\bar{x}_i$ and $\bar{u}_i$. When we have 2 periods $T = 2$.

Estimating fixed effects model: FE II

- Subtract the time-averaged equation – sometimes called the ”between equation” – from other time periods:

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

  As with the FD equation, this equation is free of a_i.

- **Remark:** We view this ”time-demeaned” (or ”within”) equation as an estimating equation. As with FD, we interpret β_1 in the levels equation as if we have estimated it in levels and controlled for a_i, as below

  $$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$$

- β_1 is called the ”fixed effects (FE) estimator” or the ”within estimator.”

Estimating fixed effects model: FE III

- There is yet another way to compute $\hat{\beta}_1$. Keep the original data, y_{it} and x_{it}, and run a regression of y_{it} on $n - 1$ dummy variables (one for each i) and x_{it}. Called the "dummy variable regression". Often not practical when n is large.

```r
## FE estimator using plm package.
reg2_fe = plm(crmrte ~ d87+unem, data=pdat, model="within")
coef(summary(reg2_fe))
```

```
##      Estimate Std. Error t-value Pr(>|t|)
## d87     15.40      4.702    3.28  0.00206
## unem     2.22      0.878    2.53  0.01519
```

Estimating fixed effects model: RE I

- What if x_{it} is (almost) constant?
  ⇒ A crucial condition in order to be able to apply FD estimator is to have enough variation in Δx_{it}. Actually, this is required by ”no perfect collinearity” of Gauss-Markov assumptions.

  ⇒ When x_{it} does not change over time (or change very little), we would expect x_{it} to be constant. Ex. gender or years of schooling for people who have completed their schooling. We will be limited in what we can learn in that case. In such cases we can not separate the effect of a_i on y_{it} from the effect of time-constant factors (such as education) and as a result we can not rely on FE or FD estimation.
- Any solution?

Estimating fixed effects model: RE II

- We already saw that when $Cov(x_{it}, a_i) = 0$ is the case, we can use POLS. But, then we said that because of the presence of a_i in the error term in each period we need to deal with serial correlation. And using heteroskedasticity-robust standard errors does not solve the problem.
- To deal with the serial correlation described above, one solution is using generalized least squares (GLS). This is known as ”random effects” (RE) estimation.
- For program evaluation, typically RE would be less convincing because one would like to participation to be correlated with the time-constant factors in a_i.
- One case where RE may be of interest is using good time-constant controls in which case more is taken out of a_i.

Estimating fixed effects model: RE III

```r
## RE estimator using plm package. Note: we use same covariates as pooling
reg2_re = plm(crmrte ~ d87+unem, data=pdat, model="random")
coef(summary(reg2_re))
```

```
##             Estimate Std. Error t-value Pr(>|t|)
## (Intercept)    80.12      9.279    8.63 2.14e-13
## d87            13.45      4.498    2.99 3.61e-03
## unem            1.75      0.811    2.16 3.37e-02
```

The coefficient on unemployment rate is now significant, which suggests that an increase in unemployment affects crime rates.