IIA-3 Econometrics: Supervision 6

Emre Usenmez

Lent Term 2025

Topics Covered

Faculty Qs:

Supplementary Qs: Independently pooled cross-section; panel data; difference-in-differences (DiD) estimator;

Related Reading:

Dougherty, *Introduction to Econometrics*, 5^{th} ed, OUP
Chapter 14: Introduction to Panel Data Models

Wooldridge J M (2021) *Introductory Econometrics: A Modern Approach*, 7^{th} ed,
Chapter 13: Pooling Cross Sections across Time: Simple Panel Data Methods

Gujarati, D N and Porter, D (2009) *Basic Econometrics*, 7^{th} International ed, McGraw-Hill
Chapter 16: Panel Data Regression Models

Gujarati, D (2022) *Essentials of Econometrics*, 5^{th} ed, Sage
Chapter 12: Panel Data Regression Models

Stock, J H and Watson M W (2020) *Introduction to Econometrics*. 4^{th} Global ed, Pearson
Chapter 10: Regression with Panel Data

Very grateful to Dr Oleg Kitov and Dr Clive Lawson for the very informative stylized answers to previous iterations of the supervision questions.

FACULTY QUESTIONS

QUESTION 1

SUPPLEMENTARY QUESTIONS

These questions are intended to guide the students through the procedures for estimation using Panel Data sets. A large part of the problem with this topic is keeping in mind the basic structure of a panel data set, there is little more theory than already covered with omitted variable bias.

Question A deals with pooled cross-sections in order to make the comparison with panel data sets in questions B and C.

QUESTION A

(1) Explain the difference between independently pooled cross section data sets and panel data sets. Is either heteroskedasticity or serial correlation likely to be more of a problem for pooled cross section estimates? Explain why.

Answer: Let's first define the two types of data sets.

Indy-pooled X-section: When we sample randomly from a population at different points in time we obtain an *independently pooled cross section*. These data sets consist of independently sampled observations which, among other things, ensures error terms across different observations would not be correlated.

One reason for using independently pooled cross sections is to increase the sample size, which would result in more precise estimators as well as in test statistics with more power. As long as the relationship between at least some of the explanatory variables and the response variable remains constant over time.

Because the data is pooled from different time periods, the populations may have different distributions in different time periods. To accomodate for this, the intercept in the model is allowed to differ across time periods. This is done by incorporating dummy variables for all but one time periods, where earliest time period is usually chosen as the base year. We can also use a time period dummy variable to check for structural changes as we did in the Michaelmas term, by interacting that dummy with a key explanatory variable of interest.

Panel Data: A *panel data*, or *longitudinal data* are different from independently pooled cross section in that the *same* unit of observation in a cross-sectional sample are surveyed across time. As a result, we cannot assume that the observations are independently distributed across time.

Given these definitions, both types of data sets can have both heteroskedasticity and serial correlation.

---

(2)

(a) Using data in the "houseprices" worksheet, and noting the definitions of each variable, estimate the following equation for 1981:

$$log(rprice)_i = \beta_0 + \beta_1 \, nearinc_i + u_i \tag{1}$$

Given that the building of the incinerator was completed before 1981, interpret your results. Do your results imply that the building of the incinerator causes house prices to fall?[1]

Answer: Since we are going to estimate this equation for only 1981 only we need to select that year in our regression. In STATA this can be done in a number of ways. The following is one of them:

```
/* load the data */
  quietly cd ..
  quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow
/* `firstrow` indicates that the first row contains the variable names */

/* run the regression */
  reg lrprice nearinc if year == 1981
```

```
      Source |       SS           df       MS      Number of obs   =       142
-------------+----------------------------------   F(1, 140)       =     38.85
       Model |    4.656479         1    4.656479   Prob > F        =    0.0000
    Residual |  16.7808989       140  .119863564   R-squared       =    0.2172
-------------+----------------------------------   Adj R-squared   =    0.2116
       Total |  21.4373779       141  .152038141   Root MSE        =    .34621

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     nearinc |  -.4025714   .0645889    -6.23   0.000    -.5302671   -.2748757
       _cons |   11.47852   .0342802   334.84   0.000     11.41074    11.54629
------------------------------------------------------------------------------
```

and in R:

```
# Load the data
houseprices_df <- read_excel("../Data/panel1.xls", sheet = "houseprices")

# create a subset of the data
houseprices_1981_df <- houseprices_df[houseprices_df$year==1981,]

# run the regression on this subset
SQA2a_lm <- lm(lrprice ~ nearinc, data = houseprices_1981_df)
summary(SQA2a_lm)
```

Since this is a simple regression on a dummy variable, the intercept of 11.47852 is the average selling price for homes that are not near the incinerator. The slope coefficient of -0.40257 for *nearinc* indicates the percentage difference in the average selling price between homes that are near the incinerator and those that are not. In this instance, houses near the incinerator seem to sell on average 0.4 percent less than those that are not near it. It is statistically significant with t value of -6.233 thus we would reject the null that there is no difference in house prices with respect to proximity to the incinerator, i.e. $\mathbb{H}_0 : \beta_1 = 0$. However, this does not imply that building incinerator caused house prices to fall, the causation could be the other way around. We can run the same regression for 1978 before the rumors about incinerator began.

---

[1] This question is from Wooldridge (2021), Section 13-2: Policy Analysis with Pooled Cross Sections, Example 13.3, and and is based on Kiel, K A and McClain, K T (1995) "House Prices During Siting Decision Stages: The Case of an Incinerator from Rumor through Operation" *Journal of Environmental Economics and Management* 28:241-255.

(b) Estimate equation (1) on page 3 using data for 1978 (at which time the incinerator had not even been planned) and calculate the difference-in-differences estimator.

Answer: We will do the same approach as we did in part (a), except this time for year 1978. This year is chosen because this is when there were not even the rumors that the incinerator would be built in North Andover, Massachusetts. Those rumors started after 1978.

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow

regress lrprice nearinc if year==1978
```

```
      Source |       SS           df       MS      Number of obs   =       179
-------------+----------------------------------   F(1, 177)       =     40.31
       Model |  4.44628835         1  4.44628835   Prob > F        =    0.0000
    Residual |  19.5249266       177   .11031032   R-squared       =    0.1855
-------------+----------------------------------   Adj R-squared   =    0.1809
       Total |   23.971215       178  .134669747   Root MSE        =    .33213

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     nearinc |  -.3399216   .0535412    -6.35   0.000    -.4455828   -.2342604
       _cons |   11.28542   .0299472   376.84   0.000     11.22632    11.34452
------------------------------------------------------------------------------
```

and in R:

```
houseprices_1978_df <- houseprices_df[houseprices_df$year==1978,]

# run the regression on this subset
SQA2b_lm <- lm(lrprice ~ nearinc, data = houseprices_1978_df)
summary(SQA2b_lm)
```

This shows that even before there was even a talk of an incinerator, the average value of a home near the site was 0.3 percent lower than the average value of a home far from the site. That difference is statistically significant with t-statistic of -6.35.This result seems to be consistent with the view that the incinerator was built in an area with lower housing values.

To see if building an incinerator impacts house prices, we can use *difference-in-differences estimator*, or (DiD estimator) which is the difference over time in the average house price differences in two locations. We can then check if this estimator is statistically different from zero.

Difference in Differences[a]

[a] Wooldridge (2021, 7^{th} ed) Section 13-2: Policy Analysis with Pooled Cross Sections

Suppose we want to understand the impact of a policy. In a natural experiment, we always have a control group that are not affected by the policy change, and a treatment group that are thought to be impacted by the policy change. To control for systematic differences between the control and treatment groups, we need two years of data, one before the policy change and one after the change. Therefore, our sample is split into four groups: the control group before the change, the control group after the change, the treatment group before the change, and the treatment group after the change. Denote the control group as a whole with C, and the treatment group with T. Create a

dummy variable dT that is equal to 1 to indicate those in the treatment group T, and 0 otherwise. Create another dummy variable $d2$ that is equal to 1 to indicate the post-policy change period, and 0 otherwise.
The equation of interest is

$$Y = \beta_0 + \delta_0\ d2 + \beta_1\ dT + \delta_1\ d2 \times dT + other\ factors$$

where δ_1 measures the effect of the policy. Without other factors in the regression, $\hat{\delta}_1$ will be the **difference-in-differences** estimator:

$$\hat{\delta}_1 = \left(\bar{Y}_{2,T} - \bar{Y}_{2,C}\right) - \left(\bar{Y}_{1,T} - \bar{Y}_{1,C}\right).$$

We can rearrange this expression as

$$\hat{\delta}_1 = \left(\bar{Y}_{2,T} - \bar{Y}_{1,T}\right) - \left(\bar{Y}_{2,C} - \bar{Y}_{1,C}\right)$$

which gives a different interpretation of the DiD estimator. The first term, $\left(\bar{Y}_{2,T} - \bar{Y}_{1,T}\right)$, is the treatment group's difference in means across the two time periods. This quantity would be a good estimator of the policy effect only if we can assume no external factors changed across the two time periods. To adjust for this possibility, we subtract from this quantity, the control group's difference in means across the two time periods. We hope that this adjustment will give a good estimator of the causal impact of the program or intervention.
Table below shows that the parameter δ_1 can be estimated in two ways:

1. compute the differences in averages between the treatment and control groups in the second time period, then compute the differences in averages between the treatment and control groups in the first time period, and finally subtract the result of the latter from the former, as in the first expression for $\hat{\delta}_1$ above.

2. compute the changes in averages over time for each of the treatment and control groups, and then difference these changes, as in the second expression for $\hat{\delta}_1$ above.

These give us a two different interpretations of $\hat{\delta}_1$.

	Before	After	After - Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment - Control	β_1	$\beta_1 + \delta_1$	δ_1

The parameter δ_1 can be given an interpretation as an *average treatment effect* (ATE) where the "treatment" is the group T in the second time period.
Finally, when include the explanatory variables, the OLS estimate of δ_1 no longer has the simple form of DiD above, but its interpretation is similar.

So in order to obtain the DiD estimate, we will do the following:

$$\hat{\delta}_1 = \left(\overline{lrprice}_{81,near} - \overline{lrprice}_{81,far}\right) - \left(\overline{lrprice}_{78,near} - \overline{lrprice}_{78,far}\right)$$

In STATA, we can do this manually by:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow

quietly gen near81 = lrprice if year==1981 & nearinc==1
quietly gen far81 = lrprice if year==1981 & nearinc==0
quietly gen near78 = lrprice if year==1978 & nearinc==1
```

```
quietly gen far78 = lrprice if year==1978 & nearinc==0

means near81 far81 near78 far78
```

```
    Variable |       Type          Obs        Mean       [95% conf. interval]
-------------+--------------------------------------------------------------
      near81 | Arithmetic           40    11.07595         10.9461   11.20579
             |  Geometric           40    11.06881        10.94137   11.19774
             |   Harmonic           40     11.0618        10.93661   11.18988
-------------+--------------------------------------------------------------
       far81 | Arithmetic          102    11.47852        11.41563    11.5414
             |  Geometric          102    11.47404        11.41077   11.53767
             |   Harmonic          102    11.46951        11.40584    11.5339
-------------+--------------------------------------------------------------
      near78 | Arithmetic           56     10.9455        10.83089   11.06011
             |  Geometric           56    10.93761        10.82696   11.04938
             |   Harmonic           56    10.93001        10.82299   11.03917
-------------+--------------------------------------------------------------
       far78 | Arithmetic          123    11.28542        11.23574   11.33511
             |  Geometric          123    11.28196        11.23173   11.33241
             |   Harmonic          123    11.27843        11.22762    11.3297
------------------------------------------------------------------------------
```

Based on the arithmetic averages, we then calculate the $\hat{\delta}_1$:

```
(11.07595-11.47852)-(10.9455-11.28542)
```

```
[1] -0.06265
```

Thus,

$$\hat{\delta}_1 = \left(\overline{lrprice}_{81,near} - \overline{lrprice}_{81,far}\right) - \left(\overline{lrprice}_{78,near} - \overline{lrprice}_{78,far}\right) = -0.06265$$

This can also be automatically obtained via `didregress` command in STATA:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow

/* Run the DiD regression */
quietly didregress (lrprice) (y81nrinc), group(nearinc) time(y81)

/* The DiD estimate is the first row of the first column of resulting table */
display r(table)[1,1]
```

```
-.0626498
```

In R it is the coefficient of the cross-product term:

```
summary(lm(lrprice ~ y81*nearinc, data=houseprices_df))
```

where the coefficient for the cross-product of $y81$ and *nearinc* gives us the same result.

---

(3) Estimate the following equations and intrepret your results:

$$log(rprice)_i = \beta_0 + \beta_1\ nearinc_i + \beta_2\ y81 \times nearinc_i + u_i \tag{2}$$

$$log(rprice)_i = \gamma_0 + \gamma_1\ y81_i + \gamma_2\ y81 \times nearinc_i + \varepsilon_i \tag{3}$$

$$log(rprice)_i = \lambda_0 + \lambda_1\ y81_i + \lambda_2\ nearinc_i + \lambda_3\ y81 \times nearinc_i + v_i \tag{4}$$

Why are the estimates so different? Use these results to test the significance of difference-in-differences estimator calculated in question A(2) above. (answers to this can be found in the help sheet, so no need to hand this in unless you have problems.)

Answer: First, lets look at the interpretation of these coefficients:

β_0: The price of houses averaged over 1978 and 1981 that are not near the incinerator.

β_1: The difference between those houses near the incinerator in 1978 and those far from the incinerator in both years, which is not that interesting.

β_2: The change in price over the two years for houses near the incinerator. This would be equivalent to the sum of λ_1 and λ_3.

γ_0: The price of houses in 1978 averaged over the whole area irrespective of their proximity to the incinerator site

γ_1: Difference between prices of houses further away from incinerator in 1981 and all houses in 1978, which is again not very interesting

γ_2: The difference in price between houses near the incinerator and those far from the incinerator in 1981. This is equivalent to the sum of λ_2 and λ_3.

λ_0: The average price of a house that is not near the incinerator in 1978

λ_1: Changes in all housing values from 1978 to 1981

λ_2: Difference in price between those houses near the incinerator and those far from it in 1978. That is, the location effect that is not due to the presence of incinerator. Recall from the earlier parts of the question above that even in 1978, homes near the incinerator site sold for less than houses not near it.

λ_3: This is the parameter of interest. It measures the decline in housing values due to the new incinerator, provided we assume that houses both near and far from the incinerator site did not appreciate at different rates for other reasons. In other words, this is the estimate of the effect of the incinerator, the difference in differences estimate. This can be understood in two ways:

as the difference between the price differences for houses near and far from the incinerator between 1978 and 1981, which is $\gamma_2 - \lambda_2$, and

as the difference between the price differences in the two years between houses that are near and far from the incinerator, which is $\beta_2 - \lambda_1$.

To estimate these we will begin with creating a new variable that is the cross product of $y81$ and $nearinc$.

in STATA:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow

generate y81nearinc = y81*nearinc
```

```
regress lrprice nearinc y81nearinc
regress lrprice y81 y81nearinc
regress lrprice y81 nearinc y81nearinc
```

```
      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(2, 318)       =     40.45
       Model |  9.76429481         2  4.88214741   Prob > F        =    0.0000
    Residual |  38.3848515       318   .12070708   R-squared       =    0.2028
-------------+----------------------------------   Adj R-squared   =    0.1978
       Total |  48.1491463       320  .150466082   Root MSE        =    .34743

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     nearinc |  -.4274575   .0518841    -8.24   0.000     -.529537    -.325378
  y81nearinc |   .1304441   .0719247     1.81   0.071    -.0110643    .2719525
       _cons |   11.37296   .0231619   491.02   0.000     11.32739    11.41853
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(2, 318)       =     28.86
       Model |  7.39703239         2   3.6985162   Prob > F        =    0.0000
    Residual |  40.7521139       318  .128151302   R-squared       =    0.1536
-------------+----------------------------------   Adj R-squared   =    0.1483
       Total |  48.1491463       320  .150466082   Root MSE        =    .35798

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         y81 |   .2994381   .0444107     6.74   0.000     .2120621    .3868141
  y81nearinc |  -.4025714   .0667845    -6.03   0.000    -.5339667   -.2711761
       _cons |   11.17908   .0267569   417.80   0.000     11.12644    11.23172
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(3, 317)       =     34.47
       Model |  11.8433207         3  3.94777358   Prob > F        =    0.0000
    Residual |  36.3058255       317  .114529418   R-squared       =    0.2460
-------------+----------------------------------   Adj R-squared   =    0.2388
       Total |  48.1491463       320  .150466082   Root MSE        =    .33842

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         y81 |   .1930939   .0453208     4.26   0.000     .1039264    .2822614
     nearinc |  -.3399216   .0545555    -6.23   0.000    -.4472582    -.232585
  y81nearinc |  -.0626498   .0834408    -0.75   0.453    -.2268176     .101518
       _cons |   11.28542   .0305145   369.84   0.000     11.22539    11.34546
------------------------------------------------------------------------------
```

Notice that the coefficient on the interaction term are all different because the first two equations suffer from omitted variable bias. The first equation leaves out $y81$ and the second equation leaves out *nearinc*. In each

case the variable left out must be related to the interaction term, so biasing the interaction term's coefficient estimate.

Also notice that the coefficient of the interaction term in the last estimation is -0.0626498 which is exactly the same as the DiD estimate above. The t-statistic for this is 4.26 which indicates that it is not significant.

Also, notice that

$$\lambda_3 = \beta_2 - \lambda_1$$
$$-0.0626498 = 0.1304441 - 0.1930939$$

$$\lambda_3 = \gamma_2 - \lambda_2$$
$$-0.0626498 = -0.4025714 + 0.3399216$$

as expected.

---

(4) Now add the following variables and comment on the differences in your results: $age, age^2, rooms, baths, log(intst), log(land)$, and $log(area)$. Comment upon the reasons for adding such variables. Explain why the coefficient for $nearinc$ is no longer significant, and why the interaction term is.

Answer: Up to now we did not take into account the characteristics of the houses. It may be that the houses selling near the incinerator may be different in 1981 than those in 1978. If that is the case, it can be important to control for such characteristics. Even if the relevant house characteristics did not change, including them can greatly reduce the error variance, which can in turn reduce the standard error of the interaction term's coefficient estimate. This is what Kiel and McClain (1995) did in their study.

Here in this question, the age of the house is controlled for using a quadratic, while also controlling for distance to the inter-state in feet ($intst$), land area in feed ($land$), house area in feet ($area$), number of rooms ($rooms$), and number of baths ($baths$).

In STATA:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow
generate y81nearinc = y81*nearinc

regress lrprice y81 nearinc y81nearinc age agesq rooms baths lintst lland larea
```

```
      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(10, 310)      =     84.91
       Model |  35.2722444         10  3.52722444   Prob > F        =    0.0000
    Residual |  12.8769019        310  .041538393   R-squared       =    0.7326
-------------+----------------------------------   Adj R-squared   =    0.7239
       Total |  48.1491463        320  .150466082   Root MSE        =    .20381
```

```
------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         y81 |   .1620725      .0285     5.69   0.000     .1059947    .2181504
     nearinc |   .0322337   .0474876     0.68   0.498    -.0612052    .1256725
  y81nearinc |  -.1315137   .0519713    -2.53   0.012    -.2337748   -.0292527
         age |  -.0083591   .0014111    -5.92   0.000    -.0111357   -.0055825
       agesq |   .0000376   8.67e-06     4.34   0.000     .0000206    .0000547
       rooms |   .0473343   .0173274     2.73   0.007       .01324    .0814285
       baths |   .0942775   .0277257     3.40   0.001     .0397232    .1488318
      lintst |  -.0614476   .0315076    -1.95   0.052    -.1234434    .0005481
       lland |   .0998448    .024491     4.08   0.000     .0516551    .1480345
       larea |   .3507718   .0514866     6.81   0.000     .2494643    .4520792
       _cons |   7.651753   .4158839    18.40   0.000     6.833441    8.470066
------------------------------------------------------------------------------
```

We see that the adjusted R-squared has risen to 0.7239. The estimate of the coefficient of the interaction term is -0.1315137 which is relatively close to that without any controls. However, its t-statistic is now -2.53 which is significant at $\alpha = 0.05$. We also see a reduction in the standard error of the interaction term estimate from 0.0834408 in the uncontrolled model to 0.0519713 in the controlled model. In the model without the full set of controls, the coefficient of the interaction term implied that because of the new incinerator, houses near it lost about 6.3% in value. However, this estimate was not statistically different from zero. Here, with the full set of controls, we see that the houses near the incinerator were devalued by about 13.2%.

Also *nearinc* is no longer significant and has a very small coefficient. This indicates that the characteristics included in this model largely capture the housing characteristics that are most important for determining housing prices.

---

(5) Test the hypothesis that the variance of the last equation (with the added variables) changes over time, i.e., that this equation suffers from the kind of heteroskedasticity we might expect to find in pooled cross-section data.

Answer: Recall from Supervisions 4 Supplementary Questions 2(d) that in order to test for violation of the homoskedasticity assumption we want to test whether u^2 is related in expected value to one or more of the explanatory variables, which, in this case, would be the variable $y81$ since we are testing whether the variance changes over time.

We can do this either manually or via STATA command 'hettest' or R function 'bptest()' from 'lmtest' package.

In STATA:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow
generate y81nearinc = y81*nearinc
quietly reg lrprice y81 nearinc y81nearinc age agesq rooms baths lintst lland larea
```

```
/* apply hettest */
  hettest y81, fstat

/* or manual calculation */
  predict u, resid
  generate u2 = u^2
  regress u2 y81
```

```
Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variable: y81

H0: Constant variance

F(1, 319) =   0.25
 Prob > F = 0.6186
```

```
      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(1, 319)       =      0.25
       Model |  .002792819         1  .002792819   Prob > F        =    0.6186
    Residual |  3.58766329       319  .011246593   R-squared       =    0.0008
-------------+----------------------------------   Adj R-squared   =   -0.0024
       Total |  3.59045611       320  .011220175   Root MSE        =    .10605

------------------------------------------------------------------------------
          u2 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         y81 |  -.0059389   .0119177    -0.50   0.619    -.0293861    .0175084
       _cons |   .0427421   .0079265     5.39   0.000     .0271472     .058337
------------------------------------------------------------------------------
```

In the manual calculation we can see that the t-statistic for $y81$ is -0.5 which means we fail to reject the null hypothesis of homoskedasticity. We can also reach the same conclusion with the F-statistic of 0.25 obtained either from the `hettest` command or from the manual approach.

---

(6) Consider the following model:

$$log(rprice)_i = \delta_0 + \delta_1\ y81_i + \delta_2\ log(dist)_i + \delta_3\ y81 \times log(dist)_i + u_i \tag{5}$$

What would you expect the sign of δ_3 to be? What does it mean if $\delta_2 > 0$? Estimate this equation and interpret your results.

Answer: δ_3 is the parameter of interest. It measures the change in housing values due to distance from the incinerator site. The interaction term is 0 if year is not 1981. For 1981, the interaction term

logarithmically increases as the distance to incinerator increases. Therefore, we would expect a positive δ_3 since it would mean that the house prices increase as you move away from the incinerator.

Similarly, if $\delta_2 > 0$, it means, holding others constant, a percentage increase in distance has a $\delta_2 > 0$ percent increase in the house prices on average.

We can estimate the equation in STATA as follows:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow

regress lrprice y81 ldist y81ldist
```

```
      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(3, 317)       =     31.39
       Model |  11.0273649         3  3.67578831   Prob > F        =    0.0000
    Residual |  37.1217814       317  .117103411   R-squared       =    0.2290
-------------+----------------------------------   Adj R-squared   =    0.2217
       Total |  48.1491463       320  .150466082   Root MSE        =     .3422

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         y81 |  -.2752139   .8050627    -0.34   0.733    -1.859155    1.308727
       ldist |   .3166879   .0515323     6.15   0.000     .2152994    .4180765
    y81ldist |   .0481864    .081793     0.59   0.556    -.1127392     .209112
       _cons |   8.058478   .5084362    15.85   0.000     7.058143    9.058814
------------------------------------------------------------------------------
```

The interaction term is positive as we expected. However, with t-statistic of 0.59 it is not significantly different from 0. Just as we did in part A(4) above, we can control for house characteristics to see if this improves.

---

(7) Add the variables listed in part A(4). What do you now conclude about the effect of the incinerator's location on house prices? How do you explain the differences between these results and those in A(4)? How might you improve your results?

Answer: We will start by adding $age, age^2, rooms, baths, log(intst), log(land)$, and $log(area)$ to the model and estimate again.

In STATA:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow

regress lrprice y81 ldist y81ldist age agesq rooms baths lintst lland larea
```

```
      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(10, 310)      =     83.07
       Model |   35.063879        10   3.5063879   Prob > F        =    0.0000
    Residual |  13.0852672       310  .042210539   R-squared       =    0.7282
-------------+----------------------------------   Adj R-squared   =    0.7195
       Total |  48.1491463       320  .150466082   Root MSE        =    .20545

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         y81 |  -.4893463   .4946922    -0.99   0.323    -1.462726    .4840328
       ldist |   .0009214   .0446168     0.02   0.984    -.0868687    .0887115
    y81ldist |   .0624666   .0502789     1.24   0.215    -.0364644    .1613976
         age |  -.0080074   .0014173    -5.65   0.000    -.0107962   -.0052187
       agesq |   .0000357   8.71e-06     4.10   0.000     .0000186    .0000528
       rooms |   .0461388   .0173442     2.66   0.008     .0120115    .0802661
       baths |   .1010486   .0278224     3.63   0.000     .0463039    .1557933
      lintst |  -.0599752   .0317218    -1.89   0.060    -.1223925     .002442
       lland |   .0953423   .0247253     3.86   0.000     .0466917    .1439929
       larea |   .3507426   .0519486     6.75   0.000     .2485261    .4529591
       _cons |   7.673864   .5015727    15.30   0.000     6.686946    8.660781
------------------------------------------------------------------------------
```

We see that the adjusted R-squared rose from 0.2217 to 0.7195 after controlling for the relevant house characteristics. The estimate of the coefficient of the interaction term is 0.0624666 which is relatively close to its value without the house characteristics controls. There is also a decline in the standard error from 0.081793 to 0.0502789 in the controlled model.The t-statistic is now 1.24 which is still not high enough to reject the null hypothesis that it is significantly different from 0. That is, the interaction term has not become significant with the addition of controls.

Notice that in Question A(4) when we ran the model with proximity to the incinerator using the dummy variable *nearinc*, the interaction term $y81 \times nearinc$ was significant. This means, for distances less than or equal to 15,840 feet between a given house and the incinerator, the model seems to work well. That is, our *dist* variable, which treats the distance as continuous, would work well for distances that are less than or equal to 15,840 feet. It appears that it does not work well, however, outside of this, given that the interaction term here is not significant.

What can we do to address this? We could replace the distance outside to be the mean distance of that outside area, and use that constant number, so number marginal effects.

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow

/* create new variable that will replace `ldist` whereby it is equal to `ldist` */
/* when near incinerator and average distance when outside */
generate dis = ldist if nearinc==1
egen meanldist0 = mean(ldist / (nearinc == 0))
replace dis = meanldist0 if nearinc == 0

/* create the new interaction term */
generate disy81=dis*y81

regress lrprice y81 dis disy81 age agesq rooms baths lintst lland larea
```

```
(225 missing values generated)
```

```
(225 real changes made)

      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(10, 310)      =     85.38
       Model |  35.3238307        10  3.53238307   Prob > F        =    0.0000
    Residual |  12.8253156       310  .041371986   R-squared       =    0.7336
-------------+----------------------------------   Adj R-squared   =    0.7250
       Total |  48.1491463       320  .150466082   Root MSE        =     .2034

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         y81 |  -.9395237   .5396602    -1.74   0.083    -2.001384    .1223364
         dis |   .0271032   .0460544     0.59   0.557    -.0635156    .1177219
      disy81 |    .107845   .0548245     1.97   0.050    -.0000303    .2157203
         age |  -.0078604   .0014164    -5.55   0.000    -.0106474   -.0050733
       agesq |   .0000349   8.71e-06     4.00   0.000     .0000177     .000052
       rooms |   .0441902   .0172266     2.57   0.011     .0102943    .0780862
       baths |   .0974965   .0275601     3.54   0.000      .043268     .151725
      lintst |  -.0715753   .0288621    -2.48   0.014    -.1283656   -.0147849
       lland |   .0969197   .0244846     3.96   0.000     .0487427    .1450967
       larea |   .3537506    .051344     6.89   0.000     .2527238    .4547775
       _cons |   7.507281   .5256345    14.28   0.000     6.473019    8.541544
------------------------------------------------------------------------------
```

We see an improvement in the t-statistic at 1.97 with p value exactly at 0.5. This is still not great, though. Perhaps we can drop *dis* which is not significant.

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow
quietly generate dis = ldist if nearinc==1
egen meanldist0 = mean(ldist / (nearinc == 0))
replace dis = meanldist0 if nearinc == 0
generate disy81=dis*y81

regress lrprice y81 disy81 age agesq rooms baths lintst lland larea
```

```
(225 real changes made)

      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(9, 311)       =     95.03
       Model |  35.3095021         9  3.92327801   Prob > F        =    0.0000
    Residual |  12.8396442       311   .04128503   R-squared       =    0.7333
-------------+----------------------------------   Adj R-squared   =    0.7256
       Total |  48.1491463       320  .150466082   Root MSE        =    .20319

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         y81 |  -1.123401   .4395479    -2.56   0.011    -1.988264   -.2585368
      disy81 |   .1266185   .0445412     2.84   0.005     .0389782    .2142588
```

```
         age |  -.0080222    .001388    -5.78   0.000    -.0107533   -.0052911
       agesq |   .0000359   8.54e-06     4.20   0.000     .0000191    .0000527
       rooms |    .045187   .0171251     2.64   0.009     .0114912    .0788828
       baths |   .0966827   .0274964     3.52   0.001     .0425802    .1507853
      lintst |  -.0665764   .0275548    -2.42   0.016    -.1207939   -.0123589
       lland |   .0983586   .0243366     4.04   0.000     .0504734    .1462438
       larea |   .3525235   .0512477     6.88   0.000     .2516875    .4533596
       _cons |   7.717446   .3852871    20.03   0.000     6.959347    8.475545
------------------------------------------------------------------------------
```

Now it looks like with the full set of controls for house characteristics, and by treating the distance as continuous only in proximity, and fixed at the average distance when far, generates a statistically significant result with t-statistics for the interaction term at 2.84. The coefficient of 0.1266185 tells us that for each percentage change in distance from the incinerator impacts the house prices on average by 12.66%, holding everything else constant.

Note that instead of averaging the distance for those distances that are far from the incinerator, we could instead have created an interaction between *ldist* and *nearinc* and interact $y81$ with that variable.

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("houseprices") firstrow
generate disinc = ldist * nearinc
generate disincy81=disinc*y81

regress lrprice y81 disincy81 age agesq rooms baths lintst lland larea
```

```
      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(9, 311)       =     94.36
       Model |   35.243305         9  3.91592278   Prob > F        =    0.0000
    Residual |  12.9058413       311  .041497882   R-squared       =    0.7320
-------------+----------------------------------   Adj R-squared   =    0.7242
       Total |  48.1491463       320  .150466082   Root MSE        =    .20371

------------------------------------------------------------------------------
     lrprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         y81 |   .1543592   .0264564     5.83   0.000     .1023031    .2064153
   disincy81 |  -.0119694    .004715    -2.54   0.012    -.0212468   -.0026921
         age |  -.0082494   .0013996    -5.89   0.000    -.0110034   -.0054954
       agesq |    .000037   8.60e-06     4.30   0.000       .00002    .0000539
       rooms |   .0458385   .0171661     2.67   0.008     .0120622    .0796148
       baths |   .0952678   .0276845     3.44   0.001     .0407952    .1497404
      lintst |  -.0707557   .0284054    -2.49   0.013    -.1266467   -.0148646
       lland |   .0989578   .0244516     4.05   0.000     .0508463    .1470693
       larea |   .3526906   .0513823     6.86   0.000     .2515897    .4537915
       _cons |   7.751441   .3893158    19.91   0.000     6.985415    8.517467
------------------------------------------------------------------------------
```

Interestingly, here we get a negative coefficient for the interaction term. This is because by multiplying *lidst* with *nearinc* we effectively make all the distances that are not near the incinerator, 0. This makes it look like the houses with higher prices are actually zero distance from the incinerator, thus it indicates that the closer you get to the incinerator, the more expensive a house becomes.

QUESTION B

In Question A we looked at a policy analysis with pooled cross sections. In this question we will look at a two-period panel data analysis.[2]

(1) Using the data from the worksheet "crime1", estimate the following relationship between *crmrte* **against** *unem* **for the year** 1987 **only:**

$$crmrte = \beta_0 + \beta_1\ unem + \varepsilon \tag{6}$$

Answer: The worksheet contains data on crime, (*crmrte*) and unemployment (*unem*) rates for 46 counties for 1982 and 1987. The *year* column consists of 82 and 87, and the dummy variable d87 is 1 if *year* is 87 and 0 otherwise. The question is effectively asking what happens if we use the 1987 cross section and run a simple regression of *crmrte* on *unem*.

Before we estimate this equation, lets first plot *crmrte* against *unem* to see the relationship between the rate of unemployment and crime rates. In STATA, you can do this by just following the instructions after clicking on the 'graphics' dropdown menu. Or you can simply type 'twoway (scatter crmrate unem)' and it should produce the scatter plot. From the scatter plot it should be clear that there is little relationship.

To estimate the equation for 1987 only, we run the following commands in STATA:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("Crime1") firstrow

regress crmrte unem if year==87
```

```
      Source |       SS           df       MS      Number of obs   =        46
-------------+----------------------------------   F(1, 44)        =      1.48
       Model |  1775.90935         1  1775.90935   Prob > F        =    0.2297
    Residual |  52674.6416        44  1197.15095   R-squared       =    0.0326
-------------+----------------------------------   Adj R-squared   =    0.0106
       Total |  54450.5509        45  1210.01224   Root MSE        =      34.6

------------------------------------------------------------------------------
      crmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        unem |  -4.161134   3.416456    -1.22   0.230    -11.04655     2.72428
       _cons |   128.3781   20.75663     6.18   0.000     86.54589    170.2103
------------------------------------------------------------------------------
```

The regression gives us

$$\widehat{crmrte} = \underset{(20.76)}{128.35} \underset{(3.42)}{-4.16\ unem}$$

[2]This question is from Wooldridge (2021), Section 13-3: Two-Period Panel Data Analysis

$$n = 46, \qquad R^2 = 0.033$$

A casual interpretation of this means that an increase in the unemployment rate *lowers* the crime rate which makes little sense. The coefficient on *unem* is not statistically significant with $|t|$-statisic of 1.22. So at best, we found no link between crime and unemployment rates.

This simple regression equation likely suffers from omitted variable problems. One possible solution is to try to control for more factors such as age distribution, gender distribution, education levels, law enforcement efforts etc in a multiple regression analysis. But many factors might be hard to control for.

Alternatively, as we discussed in Supervision 5, we can use a proxy variable. Here we suspect that our independent variable *unem* is correlated with an omitted variable but we don't know how to obtain a proxy for that omitted variable. In these types of situations, we can include, as a control, the value of the dependent variable, *crmrte* from an earlier time period. This is especially useful for policy analysis.

---

(2) Estimate equation (6) again using lagged dependent variable. Comment on your results, and the reasons for using this specification.

Answer: Some cities have had high crime rates in the past. Many of the same unobserved factors contribute to both high current crime rats and past crime rates. Inertial effects are also captured by putting in lags of the dependent variable. Using a lagged dependent variable in a cross-sectional equation provides a simple way to account for historical factors that cause *current* differences in the dependent variable which are difficult to account for in other ways.

Therefore in this question we are considering the following simple equation to explain city crime rates:

$$crmrte = \beta_0 + \beta_1\ unem + \beta_2\ crmrte_{-1} + u$$

By using the lagged dependent variable, we are partialling out those factors that affect the crime rate in both 1982 and 1987.

Before we can estimate this in STATA we need to structure the data set as a panel data set. To do that, either use the 'xtset' command or go to the 'statistics' drop down menu and select 'longitudinal/panel' then 'setup and utilities' and then 'declare data set to be panel data'. Now you have to define your cross-section and time terms. On the 'panel ID variable' dropdown box select 'County', then tick the time variable box and select $d87$. Click OK.

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("Crime1") firstrow

xtset County d87

regress crmrte unem L.crmrte
```

```
Panel variable: County (strongly balanced)
 Time variable: d87, 0 to 1
         Delta: 1 unit
```

```
      Source |       SS           df       MS      Number of obs   =        46
-------------+----------------------------------   F(2, 43)        =     38.64
       Model |  34984.0528         2  17492.0264   Prob > F        =    0.0000
    Residual |  19466.4982        43   452.70926   R-squared       =    0.6425
-------------+----------------------------------   Adj R-squared   =    0.6259
       Total |  54450.5509        45  1210.01224   Root MSE        =    21.277

------------------------------------------------------------------------------
      crmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        unem |   .9829625   2.185094     0.45   0.655    -3.423699    5.389624
             |
      crmrte |
         L1. |   1.181779   .1379824     8.56   0.000      .903511    1.460047
             |
       _cons |  -17.38646   21.27386    -0.82   0.418    -60.28929    25.51637
------------------------------------------------------------------------------
```

We see that at 0.6259 the adjusted R-squared is now much higher, and the lagged variable is significant with t-statistic of 8.56. We also see that the unemployment is now positive, but still not significant.

An alternative way to use panel data is to view the unobserved factors as consisting of two types: those that are constant and those that vary over time. We will do this next.

---

(3) It is now suggested that the relationship between crime and unemployment takes the following form:

$$crmrte_{it} = \beta_0 + \delta\ d87 + \beta_1\ unem_{it} + a_i + \varepsilon_{it} \tag{7}$$

Estimate this equation as a pooled cross section equation and comment on your results. Estimate this equation again but this time using the first difference transformation.

Answer: The variable $d87$ is 0 when $t = 1$, i.e. when year is 82, and 1 when $t = 2$, i.e. when year is 87. Therefore, the intercept for $t = 1$ is β_0, and the intercept for $t = 2$ is $\beta_0 + \delta$. Just as in using independently pooled cross-sections, allowing the intercept to change over time is important. Secular trends will cause crime rates to change over a five-year period.

The variable a_i captures all unobserved, time-constant factors that affect y_{it}. Notice that this variable does not have t subscript since it does not change over time. This variable is usually called an *unobserved effect*, or *fixed effect*, or *unobserved heterogeneity*, and the model in this question is called an *unobserved effects model* or *fixed effects model*.

In this question since i denotes different counties, we'd call a_i an *unobserved county effect* or *county fixed effect* as it represents all factors affecting county crime rates that do not change over time. These unchanging factors may, for example, be geographical features (i.e. location). a_i would also include other factors that may be roughly constant over a five-year period such as the demographic features of the population. Also different cities may have different methods for reporting crimes, and the people living in the cities might have different attitudes toward crime; these are typically slow to change.

For historical reasons cities can have very different crime rates, and historical factors are effectively captured by the unobserved effect a_i.

The error ε_{it} is often called *idiosyncratic error* or *time-varying error*, because it represents unobserved factors that change over time and affect the dependent variable.

Estimating as pooled cross section equation:

Given two years of panel data, one possible way we could try to estimate the parameter of interest, β_1, is to pool the two years and use OLS, essentially similar to Question A. This method has two drawbacks however. The most important of these two drawbacks is that in order for pooled OLS to produce a consistent estimator of β_1, we need to assume that the unobserved effect a_i is uncorrelated with $unem_{it}$.

To see this, denote $u_{it} = a_i + \varepsilon_{it}$ as the composite error. For OLS to estimate β_1 and the other parameters consistently, we assume that u_{it} is uncorrelated with $unem_{it}$ where $t = 1$, i.e. 1982, or $t = 2$, i.e. 1987. This is true irrespective of whether we use a single cross section or pool the two cross sections. Therefore, pooled OLS is biased and inconsistent if a_i and $unem_{it}$ are correlated, even if the idiosyncratic error ε_{it} is uncorrelated with $unem_{it}$. The resulting bias is sometimes called *heterogeneity bias*, but it is really just bias caused from omitting a time-constant variable.

To illustrate this, let's run our regression. Notice that since we have 46 counties and two years for each county, by pooling them we will have 92 observations.

In STATA:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("Crime1") firstrow

regress crmrte unem d87
```

```
      Source |       SS           df       MS      Number of obs   =        92
-------------+----------------------------------   F(2, 89)        =      0.55
       Model |  989.717255         2  494.858627   Prob > F        =    0.5788
    Residual |  80055.7841        89  899.503192   R-squared       =    0.0122
-------------+----------------------------------   Adj R-squared   =   -0.0100
       Total |  81045.5013        91  890.609905   Root MSE        =    29.992

------------------------------------------------------------------------------
      crmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        unem |    .426546   1.188279     0.36   0.720    -1.934539    2.787631
         d87 |   7.940413   7.975324     1.00   0.322    -7.906387    23.78721
       _cons |   93.42026   12.73947     7.33   0.000      68.1072    118.7333
------------------------------------------------------------------------------
```

The coefficient on *unem* is positive but has a very small t statistic of 0.36 and thus insignificant. So using pooled OLS on the two years has not substantially changed anything from using a single cross section. This is not surpising because using pooled OLS does not solve the omitted variables problem.

Estimating using the first difference transformation:

Usually the main reason for collecting panel data is to allow for the unobserved effect, a_i, to be correlated with the explanatory variables. Here, we want to allow the unmeasured county factors in a_i that affect the crime rate also to be correlated with the unemployment rate. To do this, we can difference the data across the two years as follows:

$$crmrte_{i2} = (\beta_0 + \delta) + \beta_1\ unem_{i2} + a_i + \varepsilon_{i2} \qquad \text{(when } t = 2\text{)}$$

$$crmrte_{i1} = \beta_0 + \beta_1\ unem_{i1} + a_i + \varepsilon_{i1} \qquad \text{(when } t = 1\text{)}$$

If we subtract the latter equation from the former then we get

$$(crmrte_{i2} - crmrte_{i1}) = \delta + \beta_1(unem_{i2} - unem_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

or

$$\Delta crmrte_i = \delta + \beta_1 \Delta unem_i + \Delta\varepsilon_i$$

Notice that the unobserved effect, a_i does not appear in this first difference transformation since it has been differenced away. Also the intercept, δ, is the *change* in the intercept from $t = 1$ to $t = 2$.

For us to be able to analyze this *first-differenced equation*, we need to assume that

- $\Delta\varepsilon_i$ and $unem_i$ are uncorrelated, which would hold if ε_{it} is uncorrelated with $unem_{it}$ in both time periods.

  This assumption might be reasonable but it can also fail. Consider for example a factor in the idiosyncratic error such as law enforcement effort. Suppose this effort increases more in counties where the unemployment rate decreases. This can cause negative correlation between $\Delta\varepsilon_i$ and $unem_i$ which would then lead to a biased estimator. This problem can be overcome to some extent by including more factors in the equation.
- $\Delta unem_i$ must have some variation across i.

  This qualification fails if *unem* does not change over time for any cross-sectional observation, or if it changes by the same amount for every observation. This is not an issue here though since *unem* changes across time for almost all cities.

  The reason why this is important is that since we allow a_i to be correlated with $unem_{it}$, we can't separate the effect of a_i on $crmrte_{it}$ from the effect of any variable that does not change over time.
- the first-differenced equation is homoskedastic. If it does not hold, we know from Supervision 4 how to test and correct for heteroskedasticity.

We can now estimate the first-differenced equation.

In STATA:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("Crime1") firstrow

quietly xtset County d87

quietly generate dcrmrte = crmrte - L.crmrte
quietly generate dunem = unem - L.unem
reg dcrmrte dunem
```

```
      Source |       SS           df       MS      Number of obs   =        46
-------------+----------------------------------   F(1, 44)        =      6.38
       Model |  2566.42989         1  2566.42989   Prob > F        =    0.0152
    Residual |  17689.5421        44  402.035047   R-squared       =    0.1267
-------------+----------------------------------   Adj R-squared   =    0.1069
       Total |   20255.972        45  450.132711   Root MSE        =    20.051

------------------------------------------------------------------------------
     dcrmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       dunem |   2.217996   .8778657     2.53   0.015     .4487743    3.987218
       _cons |   15.40219   4.702116     3.28   0.002     5.925701    24.87868
------------------------------------------------------------------------------
```

The same result can also be obtained in STATA more compactly as follows:

```
quietly xtset County d87

regress D.(crmrte unem)
```

Either approach gives us a positive, statistically significant relationship between crime rate and unemployment rate. Therefore, we can see that differencing to eliminate time-constant effects makes a big difference in this question.

The intercept tells is that even if the unemployment rate does not change between the two periods, the crime rate increases by 15.4 per 1,000 people, reflecting a secular increase in crime rates across 46 counties from 1982 to 1987.

---

(4) Verify that for t=2, the Fixed Effects transformation gives the same results as those in the previous question. Comment on your results.

Answer: The term "fixed effects" refers to the fact that each i's intercept does not vary over time, i.e. time-invariant, although the intercept may differ across different is.

Here for $t = 2$ the fixed effects transformation is:

$$crmrte_{i2} = \beta_0 + \delta\ d87 + \beta_1\ unem_{i2} + a_i + \varepsilon_{i2}$$

To be able to estimate this in STATA we need to use the 'xtreg' command with the 'fe' option as follows:

```
quietly cd ..
quietly import excel using Data/panel1.xls, sheet("Crime1") firstrow
quietly xtset County d87

xtreg crmrte d87 unem, fe
```

```
Fixed-effects (within) regression               Number of obs     =         92
Group variable: County                          Number of groups  =         46

R-squared:                                      Obs per group:
     Within  = 0.1961                                         min =          2
     Between = 0.0036                                         avg =        2.0
     Overall = 0.0067                                         max =          2

                                                F(2, 44)          =       5.37
corr(u_i, Xb) = -0.1477                         Prob > F          =     0.0082

------------------------------------------------------------------------------
      crmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
```

```
         d87 |   15.40219   4.702116     3.28   0.002     5.925701    24.87868
        unem |   2.217996   .8778657     2.53   0.015     .4487743    3.987218
       _cons |   75.40839    9.07054     8.31   0.000     57.12792    93.68887
-------------+----------------------------------------------------------------
     sigma_u |  28.529801
     sigma_e |  14.178065
         rho |  .80194675   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(45, 44) = 7.87                     Prob > F = 0.0000
```

In the output u corresponds to out a, the unobserved effect. Since it is fixed, it has no distribution and the reported $\hat{\sigma}_u$ is merely an arithmetic way to describe the range of estimated but fixed a_i. (If we were using the fixed-effects estimator of the random-effects model, this would be the estimate of σ_a assuming no omitted variables.) Similarly, e in the output is our ε in this question.

This approach also gives us a positive, statistically significant relationship between crime rate and unemployment rate.

Note: In general, if you are using the first differences estimator as in part (3), you will loose a constant term (probably best to ignore the first dummy. See Guajarti and Porter (2009) Section 16.4). If you are using fixed effects model then include all dummies in the equation, but ignore the estimated value of the constant.

— —

QUESTION C

(1) Explain what is meant by an unobserved or fixed effect, and carefully compare the differencing and fixed effects transformations used to remove these effects. How would you decide which transformation to use when: (a) $t = 2$; and (b) $t > 2$?

Answer: We have already defined unobserved or fixed effect in Question B(3). So we will look at the fixed effects model and differencing when we have two time periods as in Question B and when we have more than two time periods.

Fixed effects model when $t = 2$:[3]

Consider the model

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

where Z_i is an unobserved variable that varies from one entity i to the next but does not change over time. We can interpret this as the model having n intercepts, one for each entity. We are interested in β_1, the effect of X on Y, holding constant the unobserved entity characteristics Z.

Since we can think of the model having n intercepts, we can let $\alpha_i = \beta_0 + \beta_2 Z_i$. Then the model becomes

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

[3]Stock and Watson (2020), Sections 10.3 and 10.4; Gujarati and Porter (2009) Section 16.4; Wooldridge (2021) Sections 13.3-5, 14.1.

This is the *fixed effects regression model*, in which $\alpha_1, \ldots, \alpha_n$ are treated as unknown intercepts to be estimated, one for each entity.

This interpretation of α_i as an entity specific intercept in the model comes from the fact that the slope coefficient β_1 is the same for all the entities, but the intercept of the population regression line, α_i varies from one entity to the next.

The intercept α_i in the model can also be thought of as the "effect" of being in entity i, which means the terms $\alpha_1, \ldots, \alpha_n$ are known as *entity fixed effects*. The variation in the entity fixed effects comes from omitted variables that vary across entities but never over time, like Z_i.

Software typically computes the OLS fixed effects estimator in two steps.

Step 1: the entity-specific average is subtracted from each variable:

For the model $Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$, the average of both sides would be $barY_{it} = \alpha_i + \beta_1 \bar{X}_{it} + \bar{u}_{it}$ where $\bar{Y}_{it} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$, and $\bar{X}$ and $\bar{u}$ are defined similarly.

Therefore the model can be expressed as $Y_{it} - \bar{Y}_i = \beta_1(X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i)$, or more compactly as $\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$

Step 2: the regression is estimated using "entity-demeaned" variables. That is, β_1 can be estimated by OLS regression of the "entity-demeaned" variables $\tilde{Y}_{it}$ on $\tilde{X}_{it}$.

In fact, this estimator is identical to the OLS estimator of β_1 obtained by estimation of the fixed effects model using $T - 1$ binary variables.

We can use this approach to develop a fixed effects regression model using a binary variable. Introduce a dummy variable D_t that is equal to 0 when $t = 1$, and it is equal to 1 when $t = 2$. Here $t = 1$ corresponds to the earlier year. The dummy variable does not change across i, so it does not have the i subscript:

$$Y_{it} = \alpha_i + \delta_0 D_t + \beta_1 X_{it} + u_{it}$$

in this setup, the intercept for $t = 1$ is α_i and the intercept for $t = 2$ is $\alpha_i + \delta_0$. Thus δ_0 tells us by how much the intercept value of the of second period differs from the first period for entity i.

Fixed effects model when $t > 2$: The above reasoning can be extended to more than one time periods. In this case the general unobserved effects model is:

$$Y_{it} = \alpha_i + \delta_2\ D2_t + \cdots + \delta_T\ DT_t + \beta_1 X_{it} + u_{it}$$

for $t = 1, \ldots, T$, and where $D2_t = 1$ if $t = 2$, and 0 otherwise, $D3_t = 1$ if $t = 3$, and 0 otherwise, $\ldots$, $DT_t = 1$ if $t = T$, and 0 otherwise. Notice that here we are treating the first time period as the base, so we omit $D1_t$ and include in the model a total of $T - 1$-period dummies in addition to the intercept. Thus the intercept for $t = 1$ is α_i, for $t = 2$ it is $\alpha_i + \delta_2$, for $t = 3$, it is $\alpha_i + \delta_3$, etc. As before, we are primarily interested in β_1 and if the fixed effect α_i is correlated with any of the explanatory variables, then using pooled OLS on the $t > 2$ time periods of data results in biased and inconsistent estimates.

Differencing model when $t = 2$: Again consider the model

$$Y_{it} = \alpha_i + \delta_0 D_t + \beta_1 X_{it} + u_{it}$$

Since α_i is constant across time, we can difference the data across the two years. Specifically, for a cross-sectional observation i, write the two years as

$$\begin{aligned} Y_{i2} &= (\alpha_i + \delta_0) + \beta_1 X_{i2} + u_{i2} && \text{(when } t = 2\text{)} \\ Y_{i1} &= \alpha_i + \beta_1 X_{i1} + u_{i1} && \text{(when } t = 1\text{)} \end{aligned}$$

Subtracting the latter from the former then gives us

$$(Y_{i2} - Y_{i1}) = \delta_0 + \beta_1(X_{i2} - X_{i1}) + (u_{i2} - u_{i1})$$

or

$$\Delta Y_i = \delta_0 + \beta_1 \Delta X_i + \Delta u_i$$

where α_i is differenced away and δ_0 gives us the magnitude of change in the intercept from $t = 1$ to $t = 2$. The OLS estimator of β_1 is called the *first-differenced estimator*.

Differencing model when $t > 2$: The key assumption is that the idiosyncratic errors are uncorrelated with the explanatory variable in each time period. That is, the explanatory variables are strictly exogeneous after we take out the unobserved effect, α_i. If we have omitted an important time-varying variable, then this assumption is generally violated. Also, measurement error in one or more explanatory variables can cause this assumption to be false.

We can eliminate α_i by differencing adjacent periods. In the $T = 3$ case, we subtract $t = 1$ from $t = 2$, and $t = 2$ from $t = 3$. This gives

$$\Delta Y_i = \delta_2 \Delta D2_t + \delta_3 \Delta D3_t + \beta_1 \Delta X_{it} + \Delta U_{it}$$

for $t = 2, 3$. Notice that the expression here contains differences in the year dummies. So when $t = 2$, we have $\Delta D2_t = 1$ and $\Delta D3_t = 0$. When $t = 3$, we have $\Delta D2_t = -1$ and $\Delta D3_t = 1$.

For $t = T$ periods, this model extends to:

$$\Delta Y_i = \delta_2 \Delta D2_t + \delta_3 \Delta D3_t + \cdots + \delta_T \Delta DT_t + \beta_1 \Delta X_{it} + \Delta u_{it}$$

How to decide which transformation to use

- When $T = 2$, the fixed-effect and first-difference estimates, as well as all test statistics are *identical*, and so it does not matter which one we use. However, in this case the first-difference has the advantage of being straightforward to implement in any software package and it is easy to compute heteroskedasticity-rovust statistics after the first-differenced estimation, because when $T = 2$ the first-difference estimation is just a cross-sectional regression.
- When $T > 2$, the fixed-effect and first-difference estimators are not the same. Both are unbiased and consistent with T fixed as $N \to \infty$.
  - For large N and small T (i.e. $N > T$), the choice between fixed-effect and first-difference hinges on the relative efficiency of the estimators, and this is determined by the serial correlation in the idiosyncratic errors, u_{it}. Of course, efficiency comparisons require homoskedastic errors, so we assume homoskedasticity of the u_{it}.
    - When the u_{it} are serially uncorrelated, fixed effects is more efficient than first differencing, and the standard errors reported from fixed effects are valid. Because the unobserved effects model is typically stated with serially uncorrelated idiosyncratic errors, the fixed-effect is used more than the first-difference estimator. In many cases we can expect the unobserved factors that change over time to be serially correlated.
    - If u_{it} are serially uncorrelated with constant variance, then the correlation between Δu_{it} and Δu_{it+1} can be shown to be -0.5.
    - If u_{it} follows a stable $AR(1)$ model, then Δu_{it} will be serially correlated.
    - If u_{it} follows a random walk, meaning that there is a very substantial, positive serial correlation, then the difference Δu_{it} is serially uncorrelated, and first differencing is better. In many cases, the u_{it} exhibit some positive serial correlation, but perhaps not as much as a random walk. Then, we cannot easily compare the efficiency of the fixed-effects and first-difference estimators.
    - If there is substantial negative serial correlation in the Δu_{it}, fixed-effect is probably better.

  - When T is large, and especially wen N is not very large (i.e. $N < T$), we need to exercise caution in using the fixed-effects estimator because inference can be very sensitive to violations of the assumptions in these cases. Specifically, inference with the fixed effects estimator is potentially more sensitive to nonnormality, heteroskedasticity, and serial correlation in the idiosyncratic errors.
    * First differencing has the advantage of turning an integrated time series process into a weakly dependent process. Therefore, with first-differencing, we can appeal to the central limit theorem even in cases where $T > N$.
    * Normality in the idiosyncratic errors is not needed, and heteroskedasticity and serial correlation can be dealt with.

- When classical measurement error in one of more explanatory variables is present the the fixed effects estimator and the first difference estimator can be very sensitive. However, if each X_{itj} is uncorrelated with u_{it}, but the strict exogeneity assumption is otherwise violated, then the fixed-effect estimator likely has substantially lest bias than the first difference estimator, unless $T = 2$.

  The important theoretical fact is that the bias in the first difference estimator does not depend on T, while the bias in the fixed-effect estimator tends to 0 at the rate $1/T$.

  Strict exogeneity assumption may otherwise be violated when for example a lagged dependent variable is included among the regressors or there is a feedback between u_{it} and future outcomes of the explanatory variable.

Generally, it is difficult to choose between fixed-effect and first-difference when they give substantially different results. It makes sense to report both sets of results and to try to determine why they differ.

(2) Using the data from the "crime2" worksheet, estimate the following relationship between crime and the average sentence in 1987 and comment on your results (especially why this equation is unlikely to provide good results?):

$$lcrmrte_{87} = \beta_0 + \beta_1 lavgsen_{87} + u_{87} \tag{8}$$

Answer: This data set includes data on 90 counties in North Carolina for the years 1981 through to 1987.[4] Various factors including geographical location, attitudes toward crime, historical records, and reporting conventions might be contained in α_i.

In STATA:

```
quietly cd ..
use Data/crime4

regress lcrmrte lavgsen if d87==1
```

```
      Source |       SS           df       MS      Number of obs   =        90
-------------+----------------------------------   F(1, 88)        =      0.05
       Model |  .014695975         1  .014695975   Prob > F        =    0.8266
    Residual |   26.785002        88  .304375023   R-squared       =    0.0005
-------------+----------------------------------   Adj R-squared   =   -0.0108
       Total |   26.799698        89  .301120202   Root MSE        =     .5517

------------------------------------------------------------------------------
```

[4] This question is based on Wooldridge (2021) Example 13.9, which is based on Cornwell, C and Turnbull, W N (1994), "Estimating the Economic Model of Crime Using Panel Data", *Review of Economics and Statistics* 76:360-366.

```
     lcrmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     lavgsen |   .0458335   .2085873     0.22   0.827      -.36869    .4603569
       _cons |  -3.644001   .4690666    -7.77   0.000    -4.576173    -2.71183
------------------------------------------------------------------------------
```

The regression gives us

$$\begin{aligned} \widehat{lcrmrte} = \quad & -3.644 + 0.046\ lavgsen \\ & (0.47) \quad (0.21) \\ n = 90, \quad & R^2 = 0.0005 \end{aligned}$$

A casual interpretation of this means that a percentage increase in average sentence length *increases* the crime rate by about 0.46% which makes little sense. The coefficient on *lavgsen* is not statistically significant with a very small $|t|$-statisic of 0.22. So at best, we found no link between crime rates and average sentence length.

This simple regression equation likely suffers from omitted variable problems. One possible solution is to try to control for more factors in a multiple regression analysis. But many factors might be hard to control for.

Alternatively, as we discussed in Supervision 5, we can use a proxy variable. Here we suspect that our independent variable *lavgsen* is correlated with an omitted variable but we don't know how to obtain a proxy for that omitted variable. In these types of situations, we can include, as a control, the value of the dependent variable, *lcrmrte* from an earlier time period.

---

(3) Re-estimate equation(8) now including a lagged dependent variable in the equation (i.e. $lcrmrte_{86}$). Comment on your results and the reasoning behind including the lagged dependent variable in this case.

Answer: We will consider the following simple equation to explain county crime rates:

$$lcrmrte_{87} = \beta_0 + \beta_1 lavgsen_{87} + lcrmrte_{86} + u_{87}$$

Before we can estimate this in STATA we need to structure the data set as a panel data set, just like we did in Question B(2).

```
quietly cd ..
use Data/crime4

quietly xtset county year

regress lcrmrte lavgsen L.lcrmrte if d87==1
```

```
      Source |       SS           df       MS      Number of obs   =        90
-------------+----------------------------------   F(2, 87)        =    255.91
       Model |  22.9060927         2  11.4530463   Prob > F        =    0.0000
    Residual |  3.89360532        87  .044754084   R-squared       =    0.8547
-------------+----------------------------------   Adj R-squared   =    0.8514
       Total |   26.799698        89  .301120202   Root MSE        =    .21155

------------------------------------------------------------------------------
     lcrmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     lavgsen |  -.1540217     .08047    -1.91   0.059    -.3139647    .0059212
             |
     lcrmrte |
         L1. |   .8318358   .0367805    22.62   0.000     .7587305     .904941
             |
       _cons |  -.1877476    .236021    -0.80   0.429     -.656865    .2813697
------------------------------------------------------------------------------
```

The regression now gives us

$$\begin{aligned} \widehat{lcrmrte} = \quad & -0.188 - 0.154\ lavgsen + 0.083\ lcrmrte_{-1} \\ & (0.24) \quad (0.08) \quad (0.37) \\ n = 90, \quad & \bar{R}^2 = 0.8514 \end{aligned}$$

So we see that the adjusted R-squared is now much higher and the lagged variable is significant with a very high t-statistic of 22.62. We also see that *lavgsen* is now negative but still not significant.

---

(4) Using data for just 1987 and 1986, now estimate a pooled cross-section equation if the population relationship is thought to be of the following form:

$$lcrmrte_{it} = \beta_0 + \delta_0\ d87 + \beta_1\ lavgsen_{it} + a_i + \varepsilon_{it} \tag{9}$$

Answer: Given two years of panel data, one possible way we could try to estimate the parameter of interest, β_1, is to pool the two years and use OLS, essentially similar to Question A as well as Question B(3). This method has two drawbacks however. The most important of these two drawbacks is that in order for pooled OLS to produce a consistent estimator of β_1, we need to assume that the unobserved effect a_i is uncorrelated with $lavgsen_{it}$.

Since we have 90 counties and two years for each county, by pooling them we should have 180 observations:

```
quietly cd ..
use Data/crime4

regress lcrmrte d87 lavgsen if year>85
```

```
      Source |       SS           df       MS      Number of obs   =       180
-------------+----------------------------------   F(2, 177)       =      0.40
       Model |  .273135279         2   .13656764   Prob > F        =    0.6703
    Residual |  60.2804646       177  .340567597   R-squared       =    0.0045
-------------+----------------------------------   Adj R-squared   =   -0.0067
       Total |  60.5535999       179  .338288267   Root MSE        =    .58358

------------------------------------------------------------------------------
     lcrmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         d87 |    .074918   .0887246     0.84   0.400    -.1001762    .2500122
     lavgsen |   .0204698   .1616566     0.13   0.899    -.2985525    .3394921
       _cons |  -3.662322   .3487607   -10.50   0.000    -4.350587   -2.974058
------------------------------------------------------------------------------
```

The coefficient on *lavgsen* is positive but has a very small t-statistic of 0.84 and thus insignificant. So using pooled OLS on the two years has not substantially changed anything from using a single cross section. This is not surprising because pooled OLS does not solve the omitted variables problem.

---

(5) Show that estimating equation (ref{eq:SQC4}) using the first difference transformation (for the same years) would involve estimating the following equation:

$$\Delta lcrmrte_i = \delta + \beta_1 \Delta lavgsen_i + \Delta\varepsilon_i \tag{10}$$

Answer: Usually the main reason for collecting panel data is to allow for the unobserved effect, a_i, to be correlated with the explanatory variables. Here, we want to allow the unmeasured county factors in a_i that affect the crime rate also to be correlated with the average sentence length. To do this, denote $t = 1$ for 1986, and $t = 2$ as 1987. Then we can difference the data across the two years as follows:

$$\begin{aligned} lcrmrte_{i2} &= (\beta_0 + \delta) + \beta_1\ lavgsen_{i2} + a_i + \varepsilon_{i2} && \text{(when } t = 2\text{)} \\ crmrte_{i1} &= \beta_0 + \beta_1\ unem_{i1} + a_i + \varepsilon_{i1} && \text{(when } t = 1\text{)} \end{aligned}$$

If we subtract the latter equation from the former then we get

$$(crmrte_{i2} - crmrte_{i1}) = \delta + \beta_1(lavgsen_{i2} - lavgsen_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

or

$$\Delta crmrte_i = \delta + \beta_1 \Delta lavgsen_i + \Delta\varepsilon_i$$

as desired.

Notice that the unobserved effect, a_i does not appear in this first difference transformation since it has been differenced away. Also the intercept, δ, is the *change* in the intercept from $t = 1$ to $t = 2$.

For us to be able to analyze this *first-differenced equation*, we need to assume that

- $\Delta\varepsilon_i$ and $lavgsen_i$ are uncorrelated, which would hold if ε_{it} is uncorrelated with $lavgsen_{it}$ in both time periods.

  This assumption might be reasonable but it can also fail. Consider for example a factor in the idiosyncratic error such as law enforcement effort. Suppose this effort decreases more in counties where the sentencing length increases. This can cause negative correlation between $\Delta\varepsilon_i$ and $lavgsen_i$ which would then lead to a biased estimator. This problem can be overcome to some extent by including more factors in the equation.

- $\Delta lavgsen_i$ must have some variation across i.

  This qualification fails if $lavgsen$ does not change over time for any cross-sectional observation, or if it changes by the same amount for every observation. This is not an issue here though since $lavgsen$ changes across time for almost all counties.

  The reason why this is important is that since we allow a_i to be correlated with $lavgsen_{it}$, we can't separate the effect of a_i on $crmrte_{it}$ from the effect of any variable that does not change over time.

- the first-differenced equation is homoskedastic. If it does not hold, we know from Supervision 4 how to test and correct for heteroskedasticity.

---

(6) Again using the data for just 1987 and 1986, estimate equation (10) and verify that these results are the same as those given by using the fixed effects transformation. (You might also try to get these fixed effects results manually within Stata - i.e. not using the `xtreg` command).

Answer: Note that variable $clcrmrte$ is the first-differenced $log(crmrte)$ and $clavgsen$ is the first-differenced $log(avgsen)$. We can estimate the first-differenced equation in STATA as follows:

```
quietly cd ..
use Data/crime4

regress clcrmrte clavgsen if year > 86
```

```
      Source |       SS           df       MS      Number of obs   =        90
-------------+----------------------------------   F(1, 88)        =     14.89
       Model |  .737024111         1  .737024111   Prob > F        =    0.0002
    Residual |   4.3565703        88  .049506481   R-squared       =    0.1447
-------------+----------------------------------   Adj R-squared   =    0.1350
       Total |  5.09359441        89  .057231398   Root MSE        =     .2225

------------------------------------------------------------------------------
    clcrmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    clavgsen |  -.2827818   .0732895    -3.86   0.000    -.4284293   -.1371344
       _cons |   .1076193   .0247494     4.35   0.000      .058435    .1568035
------------------------------------------------------------------------------
```

We can also obtain the same results using either of the following commands:

```
quietly xtset county year

regress D.(lcrmrte lavgsen) if year > 86
/* OR */
xtreg lcrmrte d87 lavgsen if year > 85, fe
```

We get the same output that shows a negative and statistically significant relationship between crime rate and average sentence length. Therefore, we can see that differencing to eliminate time-constant effects makes a big difference in this question.

The intercept tells us that even if the average sentence length does not change between the two periods, the crime rate increases by about 0.1% reflecting a secular increase in crime rates across 90 counties of North Carolina from 1986 to 1987.

We can also do the fixed effects transformation manually:

```
quietly cd ..
use Data/crime4

/* generate averages */
quietly egen avc = mean(lcrmrte) if year > 85, by (county)
quietly egen avs = mean(lavgsen) if year > 85, by (county)

/* generate "entity-demeaned" variables */
quietly gen cfe = lcrmrte - avc if year > 85
quietly gen sfe = lavgsen-avs if year > 85

/* run the regression using "entity-demeaned" variables */
reg cfe d87 sfe
```

```
      Source |       SS           df       MS      Number of obs   =       180
-------------+----------------------------------   F(2, 177)       =     25.85
       Model |  .636186691         2  .318093345   Prob > F        =    0.0000
    Residual |  2.17828515       177  .012306696   R-squared       =    0.2260
-------------+----------------------------------   Adj R-squared   =    0.2173
       Total |  2.81447184       179  .015723306   Root MSE        =    .11094

------------------------------------------------------------------------------
         cfe | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         d87 |   .1076193    .017451     6.17   0.000     .0731805     .142058
         sfe |  -.2827818   .0516769    -5.47   0.000     -.384764   -.1807997
       _cons |  -.0538096    .012021    -4.48   0.000    -.0775326   -.0300867
------------------------------------------------------------------------------
```

(7) Now, using data for all years, estimate the following equation first as a pooled cross section equation, then using a first difference transformation, and finally a fixed effects transformation (N.B. for the first difference transformation, read Wooldridge on differencing with more than two time periods - Section 13.5)

$$\begin{aligned} log(crmrte)_{it} = \delta_1 + \delta_2\ d82_i + \delta_3\ d83_i + \delta_4\ d84_i + \delta_5\ d85_i + \delta_6\ d86_i + \delta_7\ d87_i + \beta_1\ log(prbarr)_{it} \\ + \beta_2\ log(prbconv)_{it} + \beta_3\ log(prbpris)_{it} + \beta_4\ log(avgsen)_{it} + \beta_5\ log(polpc)_{it} + a_i + u_{it} \end{aligned} \tag{11}$$

Answer: Pooled cross section in STATA can be obtained as follows:

```
quietly cd ..
use Data/crime4

reg lcrmrte d82 d83 d84 d85 d86 d87 lprbarr lprbconv lprbpris lavgsen lpolpc
```

```
      Source |       SS           df       MS      Number of obs   =       630
-------------+----------------------------------   F(11, 618)      =     74.49
       Model |  117.644669        11  10.6949699   Prob > F        =    0.0000
    Residual |   88.735673       618  .143585231   R-squared       =    0.5700
-------------+----------------------------------   Adj R-squared   =    0.5624
       Total |  206.380342       629  .328108652   Root MSE        =    .37893

------------------------------------------------------------------------------
     lcrmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         d82 |   .0051371    .057931     0.09   0.929    -.1086284    .1189026
         d83 |   -.043503   .0576243    -0.75   0.451    -.1566662    .0696601
         d84 |  -.1087542    .057923    -1.88   0.061     -.222504    .0049957
         d85 |  -.0780454   .0583244    -1.34   0.181    -.1925835    .0364928
         d86 |  -.0420791   .0578218    -0.73   0.467      -.15563    .0714719
         d87 |  -.0270426    .056899    -0.48   0.635    -.1387815    .0846963
     lprbarr |  -.7195033   .0367657   -19.57   0.000    -.7917042   -.6473024
    lprbconv |  -.5456589   .0263683   -20.69   0.000    -.5974413   -.4938765
    lprbpris |   .2475521   .0672268     3.68   0.000     .1155314    .3795728
     lavgsen |  -.0867575   .0579205    -1.50   0.135    -.2005023    .0269872
      lpolpc |   .3659886   .0300252    12.19   0.000     .3070248    .4249525
       _cons |  -2.082293   .2516253    -8.28   0.000    -2.576438   -1.588149
------------------------------------------------------------------------------
```

First difference transformation:

```
quietly cd ..
use Data/crime4

reg clcrmrte d83 d84 d85 d86 d87 clprbarr clprbcon clprbpri clavgsen clpolpc
```

```
      Source |       SS           df       MS      Number of obs   =       540
-------------+----------------------------------   F(10, 529)      =     40.32
       Model |   9.6004283        10   .96004283   Prob > F        =    0.0000
    Residual |  12.5963755       529  .023811674   R-squared       =    0.4325
-------------+----------------------------------   Adj R-squared   =    0.4218
```

```
       Total |  22.1968038       539  .041181454   Root MSE        =    .15431

------------------------------------------------------------------------------
    clcrmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         d83 |  -.0998658   .0238953    -4.18   0.000    -.1468071   -.0529246
         d84 |  -.0479374   .0235021    -2.04   0.042    -.0941063   -.0017686
         d85 |  -.0046111   .0234998    -0.20   0.845    -.0507756    .0415533
         d86 |   .0275143   .0241494     1.14   0.255    -.0199261    .0749548
         d87 |   .0408267   .0244153     1.67   0.095    -.0071361    .0887895
    clprbarr |  -.3274942   .0299801   -10.92   0.000    -.3863889   -.2685995
    clprbcon |  -.2381066   .0182341   -13.06   0.000    -.2739268   -.2022864
    clprbpri |  -.1650462    .025969    -6.36   0.000    -.2160613   -.1140312
    clavgsen |  -.0217607   .0220909    -0.99   0.325    -.0651574     .021636
     clpolpc |   .3984264    .026882    14.82   0.000     .3456177     .451235
       _cons |   .0077134   .0170579     0.45   0.651    -.0257961    .0412229
------------------------------------------------------------------------------
```

and fixed effect transformation

```
quietly cd ..
use Data/crime4

quietly xtset county year

xtreg lcrmrte d82 d83 d84 d85 d86 d87 lprbarr lprbconv lprbpris lavgsen lpolpc, fe
```

```
Fixed-effects (within) regression               Number of obs     =        630
Group variable: county                          Number of groups  =         90

R-squared:                                      Obs per group:
     Within  = 0.4342                                         min =          7
     Between = 0.4066                                         avg =        7.0
     Overall = 0.4042                                         max =          7

                                                F(11, 529)        =      36.91
corr(u_i, Xb) = 0.2068                          Prob > F          =     0.0000

------------------------------------------------------------------------------
     lcrmrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         d82 |   .0125802   .0215416     0.58   0.559    -.0297373    .0548977
         d83 |  -.0792813   .0213399    -3.72   0.000    -.1212027   -.0373598
         d84 |  -.1177281   .0216145    -5.45   0.000    -.1601888   -.0752673
         d85 |  -.1119561   .0218459    -5.12   0.000    -.1548716   -.0690407
         d86 |  -.0818268   .0214266    -3.82   0.000    -.1239185   -.0397352
         d87 |  -.0404704   .0210392    -1.92   0.055    -.0818011    .0008602
     lprbarr |  -.3597944   .0324192   -11.10   0.000    -.4234806   -.2961082
    lprbconv |  -.2858733   .0212173   -13.47   0.000    -.3275538   -.2441928
    lprbpris |  -.1827812   .0324611    -5.63   0.000    -.2465496   -.1190127
     lavgsen |  -.0044879   .0264471    -0.17   0.865    -.0564421    .0474663
      lpolpc |   .4241142   .0263661    16.09   0.000     .3723191    .4759093
```

```
       _cons |  -1.604135   .1685739    -9.52   0.000    -1.935292   -1.272979
-------------+----------------------------------------------------------------
     sigma_u |  .43487416
     sigma_e |  .13871215
         rho |  .90765322   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(89, 529) = 45.87                     Prob > F = 0.0000
```

---

(8) Compare and comment upon your results. In particular comment upon (a) the signs and significance of the "deterrent" variables, and (b) which results you would prefer and why (NB your answers to Question C(1)(b)).

Answer: The problem with assessing the disturbances in the pooled cross section is that the disturbances are unobservable. Most econometricians thus focus upon the first difference equation. If this is OK, then use these results, if it is not then go for fixed-effects transformation.

However, recall that we have strict exogeneity and homoskedasticity assumptions. So we need to test for these.

For heteroskedasticity test, we can use the usual Breusch-Pagan test from Supervision 4:

```
quietly cd ..
use Data/crime4
quietly reg clcrmrte d83 d84 d85 d86 d87 clprbarr clprbcon clprbpri clavgsen clpolpc

estat hettest d83 d84 d85 d86 d87 clprbarr clprbcon clprbpri clavgsen clpolpc, fstat
```

```
Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: d83 d84 d85 d86 d87 clprbarr clprbcon clprbpri clavgsen clpolpc

H0: Constant variance

F(10, 529) =   1.09
  Prob > F = 0.3655
```

With F-statistic of 1.09 we cannot reject the null hypothesis of homoskedasticity.

When we run the White test that squares the terms, then we get

```
quietly cd ..
use Data/crime4
quietly reg clcrmrte d83 d84 d85 d86 d87 clprbarr clprbcon clprbpri clavgsen clpolpc

imtest, white
```

```
White's test
H0: Homoskedasticity
Ha: Unrestricted heteroskedasticity

   chi2(50) = 257.57
Prob > chi2 = 0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df         p
---------------------+-----------------------------
  Heteroskedasticity |     257.57     50    0.0000
            Skewness |      62.96     10    0.0000
            Kurtosis |      12.82      1    0.0003
---------------------+-----------------------------
               Total |     333.35     61    0.0000
---------------------------------------------------
```

chi-square of 257.57 which means we can reject the null of homoskedasticity. However, note that if there is serial correlation, then this test is not very reliable. At best it is suggestive.

To test for serial correlation, recall that we said in Question C(1)(b) that if u_{it} is serially uncorrelated then correlation between Δu_{it} and Δu_{it+1} is -0.5.

```
quietly cd ..
use Data/crime4

quietly xtset county year

quietly reg clcrmrte d83 d84 d85 d86 d87 clprbarr clprbcon clprbpri clavgsen clpolpc

quietly predict U, r
quietly reg clcrmrte d84 d85 d86 d87 clprbarr clprbcon clprbpri clavgsen clpolpc L.U

test L.U = -0.5
```

```
 ( 1)  L.U = -.5

       F(  1,   439) =   29.79
            Prob > F =    0.0000
```

The F-statistic of 29.79 tells us that we can reject the null hypothesis $\mathbb{H}_0 : \rho = -0.5$ and conclude that serial correlation seems to be present. However, this serial correlation is not of a form that would suggest no correlation in the first equation.

In sum, we don't really know what to do. It is difficult to choose between fixed-effects and first-difference. So, report both sets of results. Alternatively, you could just go with fixed-effect since it uses more data.