Panel Data Analysis

Statistics and Econometrics

Jiahua Wu

Example 9.7. City crime rates

In this example, we have a panel data set with two periods - data on crime rates and unemployment rates were collected from a sample of 46 cities in 1982 and 1987. We want to study the impact of unemployment rates on cities' crime rates.

One straightforward approach is just to treat the sample as a cross-sectional data set, and regress *crmrte* on *unem*. We run regressions using data from 1987 only, and data from both periods.

```
# Example 9.7. City crime rates
load("crime2.RData")
crime.87 <- lm(crmrte ~ unem, data, subset = year == 87)
crime.pool <- lm(crmrte ~ unem + d87, data)
stargazer(crime.87, crime.pool, header = FALSE, type = 'latex',
          title = "Example 9.4. City Crime Rates", column.labels = c("1987","pool"))
```

Table 1: Example 9.4. City Crime Rates

	Dependent variable:	
	crmrte	
	1987	pool
	(1)	(2)
unem	−4.161	0.427
	(3.416)	(1.188)
d87		7.940
		(7.975)
Constant	128.378***	93.420***
	(20.757)	(12.739)
Observations	46	92
R^2	0.033	0.012
Adjusted R^2	0.011	−0.010
Residual Std. Error	34.600 (df = 44)	29.992 (df = 89)
F Statistic	1.483 (df = 1; 44)	0.550 (df = 2; 89)
Note:		*p<0.1; **p<0.05; ***p<0.01

The coefficients of *unem* are insignificant in both models, which suggest that there is no relationship between unemployment rates and crime rates. With this simple regression model, the result is likely biased because many relevant factors are not controlled for.

As we have a panel data set, we can control for those time invariant unobserved factor using a fixed effects panel data model. The function to estimate fixed effects model is given by *plm*() from *plm* package.

Before we discuss regressions, let us first talk about data issues. For any panel data set, we need to clearly

specify the variable for cross sectional units, and the variable indicating different time periods. We can then convert a normal data frame into a panel data frame (also from *plm* packages), where many common operations of panel data sets are properly implemented.

For this data set, we do not have a variable clearly indicating the city from which an observation was collected. Thus, we first create a *city* variable, using it as an index for the cross sectional units.

```
# create a panel data frame
data$city <- rep(1:46, each = 2)
data.p <- pdata.frame(data, index = c("city", "year"))
```

Now we are ready to estimate the fixed effects panel data model. The two common approaches include first-differenced estimation and fixed effects estimation. We need to specify $effect =$ "*individual*" (so fixed effects are included in the model), and $model =$ "fd" for first-differenced approach and $model =$ "fe" for fixed effects approach.

```
# first difference estimation
crime.fd <- plm(crmrte ~ d87 + unem, data, index = c("city", "year"),
                effect = "individual", model = "fd")

# fixed effects estimation
crime.fe <- plm(crmrte ~ d87 + unem, data, index = c("city", "year"),
                effect = "individual", model = "within")

stargazer(crime.fd, crime.fe, header = FALSE, type = 'latex', title = "Example 9.4",
          column.labels = c("fd","fe"), omit.stat = c("f", "adj.rsq"))
```

Table 2: Example 9.4

	Dependent variable:	
	crmrte	
	fd	fe
	(1)	(2)
d87		15.402***
		(4.702)
unem	−0.018	2.218**
	(0.609)	(0.878)
Observations	46	92
R^2	0.127	0.196
Note:		*p<0.1; **p<0.05; ***p<0.01

For panel data with two periods, estimations from the two approaches would be exactly the same. This is no longer the case when we have more than two periods. However, both estimators are unbiased, and differences in the estimates are typically small in most pratical projects. Model parameters need to be interpreted in the context of the original fixed effects model - see slide 12 for model interpretation. R^2 of the two models are neither informative nor comparable, as we are using different dependent variables (differences in y vs de-meaned y). The main reason of using fixed effects panel data model is because it would allow us draw more reliable causal inference by controlling for time-invariant unobserved factors.