Panel Data Analysis

BS1802 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 the regression using data from 1987 alone, and data from both periods.

```r
# 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 coefficient of *unem* is insignificant, so we find 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 regression, let us first talk about panel data manipulation. For any panel data set, we need to clearly specify the variable indicating cross sectional units, and the variable indicating time series unit. We can then convert a normal data frame into a panel data frame (also from *plm* packages), where many common operations of panel data set are properly implemented.

For this data set, we do not have a variable clearly indicating the city from which an observation is collected. Thus, we first create a *city* variable, use 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"))
```

Once we have a panel data frame, we can use the many handy functions from *plm* to analyze it. For instance, if we want to check out the index of the data set, we can use *index* function, and it would tell us the cross sectional units and time series index for all observations.

```
# index of a panel
index(data.p)
```

pdim function tells the overall structure of a panel data set, including # of cross-sectional units, # of periods for each cross-sectional unit, and total number of observations. A panel data set is called balance if each cross-sectional unit has the same number of observations.

```
# dimensions of a panel
pdim(data.p)
```

```
## Balanced Panel: n=46, T=2, N=92
```

Other common operations for panel data include taking difference of adjacent observations from the same cross-sectional unit, and extracting lagged variables. These two operations are implemented by $diff$ and *lag*, respectively.

```
head(data.p$unem, 10)
```

```
## 1-82 1-87 2-82 2-87 3-82 3-87 4-82 4-87 5-82 5-87
##  8.2  3.7  8.1  5.4  9.0  5.9 12.6  5.7 12.6  7.4
```

```
# take difference of adjacent observations
head(diff(data.p$unem), 10)
```

```
##      1-82      1-87      2-82      2-87      3-82      3-87      4-82
##        NA -4.500000        NA -2.700000        NA -3.100000        NA
##      4-87      5-82      5-87
## -6.900001        NA -5.200000
```

```
# extract lagged unemployment rate
head(lag(data.p$unem), 10)
```

```
## 1-82 1-87 2-82 2-87 3-82 3-87 4-82 4-87 5-82 5-87
##   NA  8.2   NA  8.1   NA  9.0   NA 12.6   NA 12.6
```

Now we are ready to discuss the estimation of fixed effects panel data model. The first approach is first-differenced estimation. For this approach, we need to specify $effect =$ "*individual*" (so fixed effects are included in the model), and $model =$ "fd" (using first-difference for estimation). Interpretation of estimates is discussed on slide 12.

```
# first difference estimation
crime.fd <- plm(crmrte ~ d87 + unem, data, index = c("city", "year"),
                effect = "individual", model = "fd")
stargazer(crime.fd, header = FALSE, type = 'latex', title = "Example 9.4. Fixed Effects")
```

Table 2: Example 9.4. Fixed Effects

	Dependent variable:
	crmrte
unem	2.218^{**}
	(0.878)
Constant	15.402^{***}
	(4.702)
Observations	46
R^2	0.127
Adjusted R^2	0.107
F Statistic	6.384^{**} (df = 1; 44)
Note:	*p<0.1; **p<0.05; ***p<0.01

Example 14.1. Effect of job training on firm scrap rates

In this example, we have data from 54 firms for three years 1987-1989. Some firms receive grants for training their workers in 1988 and 1989. We want to study how training grant would affect firms' scrap rates. In the fixed effects panel data model, we include two time dummies (for 1988 and 1989), a dummy indicating whether the firm receives the grant in the current year ($grant_{it}$), and a dummy indicating whether the firm receives the grant in the previous year ($grant_{i,t-1}$), as the effect of job training may well last for several years.

An alternative to first-differenced estimation is called fixed effects estimation, where we use demeaned variables in regression. Results from both first-differenced estimation and fixed effects estimation are presented in Table 3.

```
# Example 14.1. Effect of job training on firm scrap rates
load("jtrain.RData")
scrap.fe <- plm(log(scrap) ~ grant + grant_1 + d88 + d89, data,
                index = c("fcode", "year"), effect = "individual", model = "within")

scrap.fd <- plm(log(scrap) ~ 0 + grant + grant_1 + d88 + d89, data,
                index = c("fcode", "year"), effect = "individual", model = "fd")
stargazer(scrap.fe, scrap.fd, header = FALSE, type = 'latex',
          title = "Example 14.1", column.labels = c("fixed effects","first difference"))
```

With the first-differenced estimation, we need to explicitly exclude intercept in the regression to properly estimate coefficients for $d88$ and $d89$. Because after taking first difference, we have only two observations for each firm, and thus we cannot have all three of overall intercept, $d88$ and $d89$ in the model.

Results from fixed effects and first difference differ in most cases (unless we have a panel data set with only two periods). In this example, $grant$ and $d89$ are significant at 10% level in both models, however, lagged grant $grant_{-1}$ is significant at 5% level using fixed effects estimation while not significant at all using first-differenced estimation. We shall keep this in mind, when we interpret the results. R^2 is not comparable from the two models because the dependent variables are different. In the fixed effects estimation, dependent variable is demeaned $log(scrap)$, while it is the difference of $log(scrap)$ from adjacent observations using the first differenced estimation.

The relative efficiency of the two approaches depend on serial correlation in u_{it}. The function for serial correlation test is implemented with *pwartset*. The way to understand the test result is as follows. The null hypothesis is H_0 : there is no serial correlation. So we reject null, and conlude that there is serial correlation

Table 3: Example 14.1

	Dependent variable:	
	log(scrap)	
	fixed effects	first difference
	(1)	(2)
grant	−0.252*	−0.223*
	(0.151)	(0.131)
grant_1	−0.422**	−0.351
	(0.210)	(0.235)
d88	−0.080	−0.091
	(0.109)	(0.091)
d89	−0.247*	−0.277*
	(0.133)	(0.150)
Observations	162	108
R^2	0.201	0.037
Adjusted R^2	−0.237	0.009
F Statistic (df = 4; 104)	6.543***	0.985
Note:	*p<0.1; **p<0.05; ***p<0.01	

in u_{it} when p-value is sufficiently small, which is the case in this example.

```
# test for autocorrelation
pwartest(scrap.fe)
```

```
##
##  Wooldridge's test for serial correlation in FE panels
##
## data:  scrap.fe
## chisq = 52.745, p-value = 3.799e-13
## alternative hypothesis: serial correlation
```