Fixed Effects Panel Regression

R Tutorials for Applied Statistics

Note on required packages:

The following code requires the package `plm` for estimating `p`anel `l`inear `m`odels. If you have not done so on your machine, download the package `plm`. A call to `install.packages()` should only need to be done once on your computer. The call to `library()` loads the package into memory.

```
install.packages("plm")
library("plm")
```

```
## Loading required package: Formula
```

---

1 Panel Regression with Individual Fixed Effects

Consider a panel data set that has a large number of individuals, each of which is measured over at least two periods.

Consider the following regression:

$$y_{it} = \beta_0 + \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,i} + \epsilon_{it}$$
$$yit=\beta0+\alpha i+\beta1x1,it+\beta2x2,i+\epsilon it$$

Where the subscript ii denotes individuals and the subscript tt denotes time. Therefore y_{it}yit is the outcome value for individual ii and time tt. Suppose x_{it}xit is some variable of interest, and so an estimate for β_1β1 is of great interest.

Notice that the variable $x_{2,i}$x2,i does not have a tt subscript on it. This is some variable for individual ii that does not change with time. Examples of these kinds of variables common in panel data sets includes demographic information like race and sex.

The parameter α_iαi denotes any and all *time-invariant* variables for an individual ii, measurable or not, conceivable or not, that influence the value for their outcome variable. *Allowing for such a variable in the model reduces significantly the potential for omitted variable bias.*

The term α_iαi is often called the **individual fixed effect** or the **unobserved individual effect**.

The error term ϵ_{it}ϵit captures the unexplained portion of the outcome variable, and includes the effects of any non-time-invariant omitted variables.

2 Sample Estimate

The estimated sample regression model can be given by,

$$y_{it} = b_0 + a_i + b_1 x_{1,it} + b_2 x_{2,i} + e_{it}$$
yit=b0+ai+b1x1,it+b2x2,i+eit

Let's advance the time period by 1. The following sample regression model is also true:

$$y_{i,t+1} = b_0 + a_i + b_1 x_{1,i,t+1} + b_2 x_{2,i} + e_{i,t+1}$$
yi,t+1=b0+ai+b1x1,i,t+1+b2x2,i+ei,t+1

Suppose for ease of illustration that the panel covers only two period (the procedure and results that follow hold also with more sample periods)

Adding these two equations together we get:

$$y_{it} + y_{i,t+1} = 2b_0 + 2a_i + b_1(x_{1,it} + x_{1,i,t+1}) + 2b_2 x_{2,i} + e_{it} + e_{i,t+1}$$
yit+yi,t+1=2b0+2ai+b1(x1,it+x1,i,t+1)+2b2x2,i+eit+ei,t+1

If we divide by 2, then we have an expression for the sample regression equation in terms of the means for individual over the two time periods:

$$\bar{y}_i = b_0 + a_i + b_1 \bar{x}_{1,i} + b_2 x_{2,i} + 0.5(e_{it} + e_{i,t+1})$$
y¯i=b0+ai+b1x¯1,i+b2x2,i+0.5(eit+ei,t+1)

where there are no more tt subscripts because every variable that had a tt subscript has been averaged over all the periods.

Finally, subtract this equation from the original sample regression equation:

$$y_{i,t} - \bar{y}_i = b_1(x_{1,it} - \bar{x}_{1,i}) + u_i$$
yi,t−y¯i=b1(x1,it−x¯1,i)+ui

A number of terms dropped out of the regression model when we made this subtraction. The intercept b_0b0fell out and all *time-invariant* explanatory variables fell out including $x_{2,i}$x2,i and $a_i$ai. We changed notation for the residual term to u_{it}uit where $u_{it} = 0.5(e_{it} - 0.5e_{i,t+1})$uit=0.5(eit−0.5ei,t+1). If there is no correlation between residuals from the same individual over different time periods, then we can treat this new term u_{it}uit as our residual.

Let's simply the notation further. Let $\tilde{y}_{it} \equiv y_{i,t} - \bar{y}_i$y~it≡yi,t−y¯i denote the demeaned outcome variable and $\tilde{x}_{it} \equiv x_{1,it} - \bar{x}_i$x~it≡x1,it−x¯i denote the demeaned explanatory variable of interest. The regression equation becomes:

$$\tilde{y}_{it} = b_1 \tilde{x}_{1,it} + u_i$$
y~it=b1x~1,it+ui

Note that even though the time-invariant variables do not appear in this equation, they are *not* omitted variables. The time-invariant variables *are accounted for*. The estimate for the coefficient b_1b1 will measure the impact of x_1x1 on y_1y1 over and above any time-invariant factors that may influence the outcome variable. Demeaning the variables assures the effects of any time-invariant factors are differenced out.

3 Time and Individual Fixed Effects

The following is an *individual fixed effects* regression model like the above example:

$$y_{it} = \beta_0 + \alpha_i + \beta_1 x_{1,it} + \epsilon_{it},$$
yit=β0+αi+β1x1,it+ϵit,

where α_i αi is the term capturing all individual fixed effects. The time-invariant variable $x_{2,i}$ x2,i from the previous section is not explicitly included as α_i αi includes any and all time-invariant factors that may explain y_i yi.

A panel regression model may also have *individual and time fixed effects* like the following:

$$y_{it} = \beta_0 + \alpha_i + \gamma_t + \beta_1 x_{1,it} + \epsilon_{it},$$
yit=β0+αi+γt+β1x1,it+ϵit,

where γ_t γt is the *time fixed effect*. This captures the same thing a dummy variable for each time period would capture. This includes any and all variables, measurable or conceivable or not, that are *present for all individuals* in a given time period and that may influence the outcome variable. If a panel data set includes individuals from all over the United States, the time fixed effect captures any nationwide trends in the outcome variable.

Accounting for time-fixed effects and deriving an expression for an estimable regression model like in the previous section involves similar intuition and process. We will omit that here, as statistical packages like R do not require us to derive these expressions. We will only need to specify for which dimensions (individual vs time vs both) we wish to allow for fixed effects.

4 Example: Crime Rates

The code below downloads and loads into the workspace a data set on crime rates for 90 counties in North Carolina over 7 years from 1981 through 1987.

```
load(url("http://murraylax.org/datasets/crime4.RData"))
```

The data set includes the following variables of interest:

- `county`: A numerical index identifying each county
- `crmrte`: Crime rate as number of reported crimes per person
- `prbarr`: Estimated probability of being arrested when committing a crime
- `prbconv`: Estimated probability of being convicted if arrested
- `prbpris`: Estimated probability of prison time if convicted of a crime
- `avgsen`: Average prison sentence length
- `polpc`: Number of police officers per capita

4.1 Individual Fixed Effects Model

We will start by estimating the following individual fixed effects model:

$$\log(crmrte_{it}) = \beta_0 + \alpha_i + \beta_a prbarr_{it} + \beta_c prbconv_{it} + \beta_p prbpris_{it} + \beta_s avgsen_{it} + \beta_{pol} polpc_{it} + e_{it}$$
log(crmrteit)=β0+αi+βaprbarrit+βcprbconvit+βpprbprisit+βsavgsenit+βpolpolpcit+eit

We can estimate panel fixed effects model with the function `plm()`. It works much the same way as `lm()` does, but we must also tell it what dimension to use for fixed effects. There is a column in our data set called `county` which uniquely identifies each "individual" in our sample.

Here is the call to `plm()`:

```
plmcrime <- plm(log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc, data=data,
                index="county", effect="individual", model="within")
```

The parameter `index="county"` told `plm()` to use fixed effects based on the county variable. The parameter `effect="individual"` indicated that this is an *individual* fixed effects model. The parameter `model="within"` told `plm()` to use the fixed effects model (there are many other estimation methods for panel models that is built into `plm()`).

We can call a summary for the output as we usually do:

```
summary(plmcrime)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(crmrte) ~ prbarr + prbconv + prbpris + avgsen +
##     polpc, data = data, effect = "individual", model = "within",
##     index = "county")
##
## Balanced Panel: n = 90, T = 7, N = 630
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.7882633 -0.0817435 -0.0033513  0.0806637  0.6423049
##
## Coefficients:
##           Estimate Std. Error t-value  Pr(>|t|)
## prbarr  -0.3684527  0.0663990 -5.5491 4.524e-08 ***
## prbconv -0.0280621  0.0052219 -5.3739 1.152e-07 ***
## prbpris -0.1768960  0.0963986 -1.8350  0.067054 .
## avgsen   0.0083636  0.0030382  2.7528  0.006109 **
## polpc   42.7467994  4.2984648  9.9447 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    17.991
## Residual Sum of Squares: 14.734
## R-Squared:      0.18102
## Adj. R-Squared: 0.037129
## F-statistic: 23.6509 on 5 and 535 DF, p-value: < 2.22e-16
```

We can see that there is no intercept estimated in the model. We demonstrated in Section 2 that the intercept is differenced out along with any individual fixed effects.

We see the expected signs on several of the coefficients. The higher are the probabilities for arrest, conviction, and prison sentence, the lower is the crime rate.

The signs on average sentence length and police per capita maybe unexpected. They are both positive. It would be tempting, but probably wrong, to say that more police per capita leads to more crime. It could be, rather, that more crime gets reported when there is a greater police presence. Also, it could be that when greater crime numbers were expected, there was a larger police force. Note, however, that county fixed effects are accounted for.

We *cannot* say that the positive coefficient is due to cities with higher average crime rates in the first place hire a greater police force. If throughout the sample, there is a city that has a higher average crime rate for whatever reason, and this led to more police officers, then this is accounted for in the fixed effect.

4.2 Individual and Time Fixed Effects Model

Let us now consider the following individual and time fixed effects model:

$$\log(crmrte_{it}) = \beta_0 + \alpha_i + \gamma_t + \beta_a prbarr_{it} + \beta_c prbconv_{it} + \beta_p prbpris_{it} + \beta_s avgsen_{it} + \beta_{pol} polpc_{it} + e_{it}$$

log(crmrteit)=β0+αi+γt+βaprbarrit+βcprbconvit+βpprbprisit+βsavgsenit+βpolpolpcit+eit

Note the presence of γ_tγt, the time fixed effect. There is a column in our data set called `year` which identifies the years to used for the time fixed effect. Again present in our equation is α_i αi the county-level (individual) fixed effect. We estimate the model with the following calls to `plm()`:

```
plmcrime <- plm(log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + polpc, data=data,
                index=c("county", "year"), effect="twoways", model="within")
```

We set the parameter `index` to a list of variable names so that we can have fixed effects based on `county` and on `year`. The function `c()` combines these names into a list to pass as the index parameter. The parameter `effect="twoways"` indicates that this is a two-way fixed effects model, i.e. including time and individual fixed effects.

Let us look at a summary of the regression:

```
summary(plmcrime)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = log(crmrte) ~ prbarr + prbconv + prbpris + avgsen +
##     polpc, data = data, effect = "twoways", model = "within",
##     index = c("county", "year"))
##
## Balanced Panel: n = 90, T = 7, N = 630
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.7035582 -0.0711036  0.0036359  0.0687824  0.6809770
##
## Coefficients:
##           Estimate Std. Error t-value  Pr(>|t|)
## prbarr  -0.3459067  0.0626639 -5.5200 5.317e-08 ***
## prbconv -0.0260086  0.0049276 -5.2782 1.907e-07 ***
## prbpris -0.1247933  0.0922299 -1.3531    0.1766
## avgsen   0.0020541  0.0030580  0.6717    0.5021
## polpc   45.1199470  4.0409795 11.1656 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    16.123
## Residual Sum of Squares: 12.782
## R-Squared:      0.20722
## Adj. R-Squared: 0.057353
## F-statistic: 27.6539 on 5 and 529 DF, p-value: < 2.22e-16
```

After accounting for time- and individual-fixed effects, we do not have sufficient statistical evidence that probability of prison time and average prison sentence explain crime rate. Still, we do have sufficient statistical evidence that the probability of being arrested and the probability of being convicted both do lead to lower crime rates.