EXCERSISES IN APPLIED PANEL DATA ANALYSIS #1

CHRISTOPHER F. PARMETER

"There are two ways to write error-free programs; only the third one works." - Alan J. Perlis

1. Introduction

This set of exercises are designed to help you gain familiarity with the `plm` package and to find comfort in working with data inside of `R`. The habits and skills we develop with this exercise will be invaluable for later computer exercises.

2. The `plm` Package

The `plm` package (Croissant & Millo 2008) is the main set of calls for conducting panel data analysis in `R`. This package allows the user to access a variety of estimators and tests designed specifically for panel data. The current exercises will focus on pooled esitmation of the linear panel data model.

3. Getting the `plm` Package

To install the `plm` package on your machine (assuming you have installed `R`) simply type:

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

Once you have installed the package on your machine you can access it from the console by typing:

```
> ## Load plm package
> library(plm)
```

To access the help files within the `plm` package simply call

```
> ?plm
```

or pull up the vignette via

```
> vignette("plm")
```

University of Miami
Date: August 14, 2013.
Christopher F. Parmeter, Department of Economics, University of Miami; e-mail: cparmeter@bus.miami.edu.
*These exercises have been prepared for the "Applied Panel Data Econometrics" course in Dakar, Senegal sponsored by IFPRI and AGRODEP..

4. An Overview of the plm Package

There are four different estimation routines within the `plm package`:

- `plm`, which is the main function we will deploy throughout the exercises. This function will allow you to estimate the pooled, fixed and random effects model as well as implmenting instrumental variables routines
- `pvcm`, which estimates panel data models with random coefficients
- `pgmm`, which can be used to estimate dynamic panel data models
- `pggls`, which will allow estimation via feasible generalized least squares.

Following the introduction to `R` on the course webpage the use of these four functions is through a formula interface. While the four functions all allow one to pass a `data.frame()` to the function, it is easier to just use the `pdata.frame()` setup. The `pdata.frame()` will explicitly account for the panel nature of the data, through the user *explicitly* informing which index corresponds to the individual and which index corresponds to time. Let's try this out. First we will load in our dataset for this exercise. The dataset is a panel on 54 firms covering the years 1987-1989 and are taken from Holzer, Block, Cheatham & Knott (1993). The data is named `jtrain.csv`.

We load the data as

```
> data <- read.csv(file="jtrain.csv",h=TRUE,na.strings=".")
```

To construct a `pdata.frame()` where the time index is `Year` and the firm index is `Fcode` we type

```
> pdata <- pdata.frame(data,index=c("Fcode","Year"))
```

Now to estimate the pooled panel data model for the following specification:

$$\log(scrap_{it}) = \beta_0 + \beta_1 D_{88} + \beta_2 D_{89} + \beta_3 grant_{it} + \beta_4 grant_{it-1} + \varepsilon_{it}. \tag{1}$$

This model looks at the impact that state grants given to local firms had on scrap rates. Grants were given out starting in 1988. This is important because it allows us to know that when we have lagged grant, $grant_{it-1}$ there are no missing values.

```
> ## First construct logarithm of scrap rate
> pdata$lscrap <- log(pdata$scrap)
> ## Create lag of grant
> pdata$grant.1 <- lag(pdata$grant,1)
> ## Replace NAs with 0
> cut <- seq(from=1,to=length(pdata$grant),by=3)
> pdata$grant.1[cut] <- 0
```

Now we can estimated the pooled linear panel data model.

```
> ## Estimate pooled model
> model.pooled.scrap <- plm(lscrap~d88+d89+grant+grant.1,data=pdata,model="pooling")
```

To construct the standard errors for the pooled model simply store the `summary()` of `model.pooled.scrap`

```
> sum.pooled.scrap <- summary(model.pooled.scrap)
> ## Print out results to screen
> sum.pooled.scrap

Oneway (individual) effect Pooling Model

Call:
plm(formula = lscrap ~ d88 + d89 + grant + grant.1, data = pdata,
    model = "pooling")

Balanced Panel: n=54, T=3, N=162

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-5.2000 -0.8960 -0.0846  1.0200  3.3000

Coefficients :
             Estimate Std. Error t-value Pr(>|t|)
(Intercept)  0.597434   0.203063  2.9421 0.003754 **
d88         -0.239370   0.310864 -0.7700 0.442447
d89         -0.496524   0.337928 -1.4693 0.143748
grant        0.200020   0.338285  0.5913 0.555186
grant.1      0.048936   0.436066  0.1122 0.910792
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    355.75
Residual Sum of Squares: 349.59
R-Squared      :  0.017311
      Adj. R-Squared :  0.016777
F-statistic: 0.691427 on 4 and 157 DF, p-value: 0.59893
```

Notice that both the current and lagged effects of the grant have the wrong sign, one would assume that a state grant should help to lower the scrap rate, and neither effect is statistically significant. Notice also the low R^2. This probably suggests that a firm's scrap rate is a more complex statistical object than the model we have used (which is admittedly simple). Lets see what would happen if we included the lagged scrap rate. Given that we will lose the year 1987 by including the lagged value of the logarithm of the scrap rate, we have to drop one of our year dummies.

```
> pdata$lscrap.1 <- lag(pdata$lscrap,1)
> ## Estimate pooled model
```

```
> model.pooled.scrap1 <- plm(lscrap~d88+grant+grant.1+lscrap.1,data=pdata,model="pooling")
> sum.pooled.scrap1 <- summary(model.pooled.scrap1)
> ## Print out results to screen
> sum.pooled.scrap1


Oneway (individual) effect Pooling Model

Call:
plm(formula = lscrap ~ d88 + grant + grant.1 + lscrap.1, data = pdata,
    model = "pooling")

Balanced Panel: n=54, T=2, N=108

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-2.8800 -0.1250  0.0716  0.2490  1.8900

Coefficients :
             Estimate Std. Error t-value Pr(>|t|)
(Intercept) -0.152525   0.100205 -1.5221   0.1310
d88          0.115389   0.119913  0.9623   0.3382
grant       -0.172392   0.125744 -1.3710   0.1734
grant.1     -0.107323   0.161038 -0.6664   0.5066
lscrap.1     0.880822   0.035796 24.6065   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    217.61
Residual Sum of Squares: 31.23
R-Squared      :  0.85649
      Adj. R-Squared :  0.81683
F-statistic: 153.675 on 4 and 103 DF, p-value: < 2.22e-16
```

Now the grant effects have the correct sign, but they are still statistically insignificant. Further, the R^2 has significantly increased from the model which omitted the dynamics in scrap rate.

Let's consider another example. This example follows from Ram (2009), who looks at the debate over the relationshp between how open a country is to international trade, the size of the country (as measured by popoulation) and the amount of money that the country's government spends.

We can specify these individual relationships as:

$$\log(gov_{it}) = \alpha_0 + \alpha_1 \log(open_{it}) + \alpha_2 \log(GDP_{it}) + \varepsilon_{1,it} \quad (2)$$

$$\log(open_{it}) = \beta_0 + \beta_1 \log(size_{it}) + \beta_2 \log(GDP_{it}) + \varepsilon_{2,it} \quad (3)$$

$$\log(gov_{it}) = \gamma_0 + \gamma_1 \log(size_{it}) + \gamma_2 \log(GDP_{it}) + \varepsilon_{3,it} \quad (4)$$

where *gov* measures government spending (a proxy for government size), *open* measures the ratio of total imports and exports to GDP (a proxy for openness to trade), *size* measures the population of a country (a proxy for the size of the country), and *GDP* is the gross domestic product per capita in the country. Rodrik (1998) argues for the specification of government spending in (2) while Alesina & Wacziarg (1998), argues that because of the relationship in (3), the appropriate government spending relationship is that in (4).

We use the `pwt` package to construct the dataset used by Ram (2009) to adjudicate between these two alternative propositions for government spending. We first focus on insights from the pooled model.

```
> ## install pwt package
> install.packages("pwt")
> ##load pwt library
> library(pwt)
> ## First replicate based on PWT6.1
> data("pwt6.1")
> ## Create data.frame
> data6.1<- data.frame(pwt6.1)
> #Subset of years 1960-2000, to match Ram
> data6.1.ram <- data.frame(subset(data6.1,data6.1$year>1959 & data6.1$year<2001))
> ## First see which countries only have a single observation over the entire period
> ## Ram has 154 countries, but we have 168 countries
> uni.country <- unique(data6.1.ram$country)
> length(uni.country)
[1] 168
> ## Now keep only pop, rgdpch and openk (Ram uses openc -- current prices)
> ## But rgdpch is in chain prices !!#$#@
> check.ram <- na.omit(data6.1.ram[,c(1:4,10,15,20,24)])
> ## Which countries have a single observation
> singleyear.store <- as.numeric()
> for (i in 1:length(uni.country)){
+
+         c.id <- uni.country[i]
+
```

```
+         singleyear.store[i] <- length(which(check.ram$country==c.id))
+
+ }
> num.single <- length(which(singleyear.store==1))
> ## Remove the single year countries
> if(num.single>0){
+
+         ram.omit <- uni.country[which(singleyear.store==1)]
+
+         id.omit <- which(check.ram$country%in%ram.omit)
+         check.ram <- check.ram[-id.omit,]
+
+ }
> ## Subset of years 1960-2000, to match Ram
> pdata6.1.ram <- pdata.frame(check.ram,index=c("country","year"))
> ## Estimate Pooled OLS w/ Heteroscedasticity robust standard errors
> ## Table 1, Column (1)
> model1.1 <- plm(log(cg)~log(pop)+log(rgdpch),
+                                 model="pool",
+                                 data=pdata6.1.ram)
> robust.t.1.1 <- coefficients(model1.1)/sqrt(diag(vcovHC(model1.1,type="HC0",
+
+
> coefficients(model1.1)

(Intercept)    log(pop) log(rgdpch)
 4.65282789 -0.07951004 -0.13894406

> robust.t.1.1

(Intercept)    log(pop) log(rgdpch)
   62.25405   -17.43125   -18.83540

> ## Table 2, Column (1)
> model2.1 <- plm(log(openc)~log(pop)+log(rgdpch),
+                                 model="pool",
+                                 data=pdata6.1.ram)
> robust.t.2.1 <- coefficients(model2.1)/sqrt(diag(vcovHC(model2.1,type="HC0",
+
+
> coefficients(model2.1)
```

```
(Intercept)    log(pop) log(rgdpch)
  4.5140136  -0.2048902   0.1584018
> robust.t.2.1
(Intercept)    log(pop) log(rgdpch)
   71.28379   -55.69919    24.29159
> ## Table 3, Column (1)
> model3.1 <- plm(log(cg)~log(openc)+log(rgdpch),
+                                  model="pool",
+                                  data=pdata6.1.ram)
> robust.t.3.1 <- coefficients(model3.1)/sqrt(diag(vcovHC(model3.1,type="HC0",
+
+
> coefficients(model3.1)
(Intercept)  log(openc) log(rgdpch)
  3.2816535   0.2443707  -0.1763147
> robust.t.3.1
(Intercept)  log(openc) log(rgdpch)
   42.90547    17.08322   -23.53346
>
```

We see from the results of `model2.1` that the model in (3) supports Alesina & Wacziarg's (1998) position that Rodrik's (1998) finding is misconstrued given the link between trade openness of a country and the size of the populace. When we investigate these models in more detail using models that exploit the panel structure we will have more to say.

5. Conclusion

These exercises were designed to provide you with basic familiarity with estimating the linear pandel data model using the pooled OLS estimator. We took two different examples from the published applied economics literature and found interesting insights from both.

References

Alesina, A. & Wacziarg, R. (1998), 'Openness, country size, and government size', *Journal of Public Economics* **69**, 305–321.

Croissant, Y. & Millo, G. (2008), 'Panel data econometrics in R: The plm package', *Journal of Statistical Software* **27**(2).
URL: *http://www.jstatsoft.org/v27/i02/*

Holzer, H. M., Block, R., Cheatham, M. & Knott, J. (1993), 'Are training subsidies effective? The Michigan experience', *Industrial and Labor Relations Review* **46**, 625–636.

Ram, R. (2009), 'Openness, country size, and government size: Additional evidence from a large cross-country panel', *Journal of Public Economics* **93**, 213–218.

Rodrik, D. (1998), 'Why do more open economies have bigger governments?', *Journal of Political Economy* **106**, 997–1032.