Panel Data

Jayashree Raman

9/18/2018

```
library(plm)
```

```
## Warning: package 'plm' was built under R version 3.5.1
```

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

```
library(prediction)
```

```
## Warning: package 'prediction' was built under R version 3.5.1
```

```
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 3.5.1
```

```
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 3.5.1
```

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.5.1
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
train_data <- na.omit(read.csv(file="C:/Users/jayashree.raman/Documents/Learning/MIDS/capstone/datasets/

train_data$session_start <- as.POSIXct(train_data$session_start, "%Y-%m-%d %H:%M:%S")
```

```
## Warning in strptime(xx, f, tz = tz): unknown timezone '%Y-%m-%d %H:%M:%S'

## Warning in as.POSIXct.POSIXlt(x): unknown timezone '%Y-%m-%d %H:%M:%S'

## Warning in strptime(x, f, tz = tz): unknown timezone '%Y-%m-%d %H:%M:%S'

## Warning in as.POSIXct.POSIXlt(as.POSIXlt(x, tz, ...), tz, ...): unknown
## timezone '%Y-%m-%d %H:%M:%S'
```

```
train_data$id <- as.numeric(str_replace_all(train_data$userid, "user_", ""))
panel.data.train <- plm.data(train_data, index = c("id", "session_start"))
```

```
## Warning: use of 'plm.data' is discouraged, better use 'pdata.frame' instead

## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%m-%d %H:%M:%S'

## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%m-%d %H:%M:%S'
```

```
mdl_pooled <-plm(session_length~age+session_length_mvavg+previous_duration+absence_time+is_holiday, data

mdl_random <-plm(session_length~age+session_length_mvavg+previous_duration+absence_time+is_holiday, data

mdl_fe <-plm(session_length~age+session_length_mvavg+previous_duration+absence_time+is_holiday, data = p

##Summaries
stargazer(mdl_pooled, mdl_fe, mdl_random,
          column.labels = c("Pooled OLS", "Fixed Effects", "Random Effects"),
          type='latex',
          title="Comparing Pooled OLS, Fixed and Random Effects",
          omit="as.factor",
          covariate.labels = c("Age", "Session Length Moving Average"),
          notes = c("Fixed effects estimated but not shown in Fixed Effects column"),
          add.lines = list(c("Fixed Effects?", "No", "Yes")))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Nov 01, 2018 - 6:58:36 PM

Table 1: Comparing Pooled OLS, Fixed and Random Effects

	Dependent variable:		
	session_length		
	Pooled OLS	Fixed Effects	Random Effect
	(1)	(2)	(3)
Age	0.310	−154.020***	0.310
	(3.590)	(24.207)	(3.590)
Session Length Moving Average	0.695***	0.655***	0.695***
	(0.006)	(0.017)	(0.006)
previous_duration	0.160***	0.145***	0.160***
	(0.002)	(0.002)	(0.002)
absence_time	−0.00001	−0.00001	−0.00001
	(0.00005)	(0.00005)	(0.00005)
is_holiday	−26.509	−3.890	−26.509
	(51.160)	(58.847)	(51.160)
Constant	638.549***		638.549***
	(101.024)		(101.024)
Fixed Effects?	No	Yes	
Observations	169,556	169,556	169,556
R^2	0.123	0.036	0.123
Adjusted R^2	0.123	0.035	0.123
F Statistic	4,760.300*** (df = 5; 169550)	1,263.973*** (df = 5; 169377)	4,760.300*** (df = 5;

Note: *p<0.1; **p<0.05; ***
Fixed effects estimated but not shown in Fixed Effects

Tests

```
##Hausman test - To decide between the fixed and random effects model. Null hypothesis is that Random e

phtest(mdl_fe, mdl_random)
```

```
Hausman Test
```

data: session_length ~ age + session_length_mvavg + previous_duration + ... chisq = 9497.4, df = 5, p-value < 2.2e-16 alternative hypothesis: one model is inconsistent

```
# Breusch-Pagan test - Testing between random effects regression and a fixed effect regression - Null i
plmtest(mdl_fe, c("session_start"), effect = c("twoways"), type=("bp"))
```

```
Lagrange Multiplier Test - two-ways effects (Breusch-Pagan) for
unbalanced panels
```

data: session_length ~ age + session_length_mvavg + previous_duration + ... chisq = 11975, df = 2, p-value < 2.2e-16 alternative hypothesis: significant effects

```
# Breusch-Pagan test - Testing between random effects regression and a simple OLS regression - Null is
plmtest(mdl_pooled, effect = c("twoways"), type=("bp"))
```

```
Lagrange Multiplier Test - two-ways effects (Breusch-Pagan) for
unbalanced panels
```

data: session_length ~ age + session_length_mvavg + previous_duration + ... chisq = 11975, df = 2, p-value < 2.2e-16 alternative hypothesis: significant effects

```
# Breusch--Godfrey Test For Panel Models - Test of serial correlation for (the idiosyncratic component
pbgtest(mdl_fe)
```

```
Breusch-Godfrey/Wooldridge test for serial correlation in panel
models
```

data: session_length ~ age + session_length_mvavg + previous_duration + absence_time + is_holiday chisq = 11295, df = 19, p-value < 2.2e-16 alternative hypothesis: serial correlation in idiosyncratic errors

```
# F Test For Individual And/Or Time Effects- Test of individual and/or time effects based on the compar

pFtest(mdl_fe, mdl_pooled)
```

```
F test for individual effects
```

data: session_length ~ age + session_length_mvavg + previous_duration + ... F = 14.383, df1 = 173, df2 = 169380, p-value < 2.2e-16 alternative hypothesis: significant effects

```r
# The Dickey-Fuller test to check for stochastic trends. The null hypothesis is that the series has a u

adf.test(panel.data.train$session_length, k=3)
```

```
## Warning in adf.test(panel.data.train$session_length, k = 3): p-value
## smaller than printed p-value
```

```
Augmented Dickey-Fuller Test
```

data: panel.data.train$session_length Dickey-Fuller = -128.7, Lag order = 3, p-value = 0.01 alternative hypothesis: stationary

```r
#Panel Unit root tests
purtest(session_length ~ trend, data=train_data, index=c('session_start', 'id'), pmax=8, exo = "trend",
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%m-%d %H:%M:%S'
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%m-%d %H:%M:%S'
```

```
Hadri Test (ex. var.: Individual Intercepts and Trend)
(Heterosked. Consistent)
```

data: session_length ~ trend z = NaN, p-value = NA alternative hypothesis: at least one series has a unit root