Panel Data

Jayashree Raman

9/18/2018

```
library(car)
```

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

```
## Loading required package: carData
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.1
```

```
library(sqldf)
```

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

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
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
```

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:

```r
usrSessData <- read.table(file="usersessions-jayashree.csv", header=TRUE,
row.names = NULL, sep=",")
SessData<- na.omit(usrSessData)

str(SessData)
```

```
## 'data.frame':    135879 obs. of  10 variables:
##  $ userid        : Factor w/ 564 levels "user_000001",..: 3 3 3 3 3 3 3 3
3 3 ...
##  $ gender        : Factor w/ 3 levels "","f","m": 2 2 2 2 2 2 2 2 2 2 ...
##  $ age           : int  19 19 19 19 19 19 19 19 19 19 ...
##  $ country       : Factor w/ 56 levels "","Antarctica",..: 53 53 53 53 53
53 53 53 53 53 ...
##  $ startdate     : Factor w/ 1586 levels "2005-02-14","2005-02-15",..: 698
699 699 699 699 700 700 700 700 700 ...
##  $ day_of_week   : Factor w/ 7 levels "friday   ","monday   ",..: 3 4 4 4
4 2 2 2 2 2 ...
##  $ timeofday     : Factor w/ 4 levels "evening","morning",..: 1 3 2 2 1 3
3 2 4 4 ...
##  $ sessionid     : int  1 1 2 3 4 1 2 3 4 5 ...
##  $ session_start : Factor w/ 412743 levels "2005-02-14 00:00:07",..:
106050 106178 106187 106190 106438 106498 106502 106513 106606 106645 ...
##  $ session_length: int  5 57 14 149 65 8 3 42 30 33 ...
##  - attr(*, "na.action")= 'omit' Named int  1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr  "1" "2" "3" "4" ...
```

```r
par(mfrow=c(2, 2))
table(SessData$country)
```

```
##
##
##                              2935
##                        Antarctica
##                                 0
##                         Argentina
##                               227
##                           Armenia
##                              1865
##                         Australia
##                              5215
##                           Austria
##                                 0
##                           Belgium
##                              2890
##               Bosnia and Herzegovina
##                                 0
##                            Brazil
##                              2758
##                          Bulgaria
##                                 0
```

```
##                                  Canada
##                                    4442
##                                   Chile
##                                     189
##                                   China
##                                       0
## Congo, the Democratic Republic of the
##                                       0
##                                 Croatia
##                                     992
##                           Czech Republic
##                                    2054
##                                 Estonia
##                                     896
##                                 Finland
##                                    5020
##                                  France
##                                    1012
##                                 Germany
##                                    2917
##                                  Greece
##                                     782
##                                 Hungary
##                                    1491
##                                 Ireland
##                                       0
##                                  Israel
##                                       0
##                                   Italy
##                                    9768
##                                   Japan
##                                       4
##                                  Latvia
##                                       0
##                               Lithuania
##                                       0
##                               Macedonia
##                                       0
##                                   Malta
##                                       0
##                                  Mexico
##                                    1778
##                                 Morocco
##                                     841
##                             Netherlands
##                                    1396
##                    Netherlands Antilles
##                                       0
##                             New Zealand
##                                     558
```

```
##                               Nicaragua
##                                     111
##               Northern Mariana Islands
##                                       0
##                                  Norway
##                                    3353
##                                    Peru
##                                     183
##                                  Poland
##                                   13826
##                                Portugal
##                                       0
##                                 Romania
##                                    2214
##                      Russian Federation
##                                    4876
##                                  Serbia
##                                     941
##                                Slovakia
##                                       0
##                                   Spain
##                                    4096
##                                  Sweden
##                                    6980
##                             Switzerland
##                                       0
##                                Thailand
##                                    1188
##                     Trinidad and Tobago
##                                       0
##                                  Turkey
##                                    9992
##                           United Kingdom
##                                   14468
##                            United States
##                                   18019
##  United States Minor Outlying Islands
##                                    1979
##                                Venezuela
##                                    3623
##                                Zimbabwe
##                                       0
```

```
hist(SessData$age)
table(SessData$gender)
```

```
##
##           f     m
##  2884 55740 77255
```

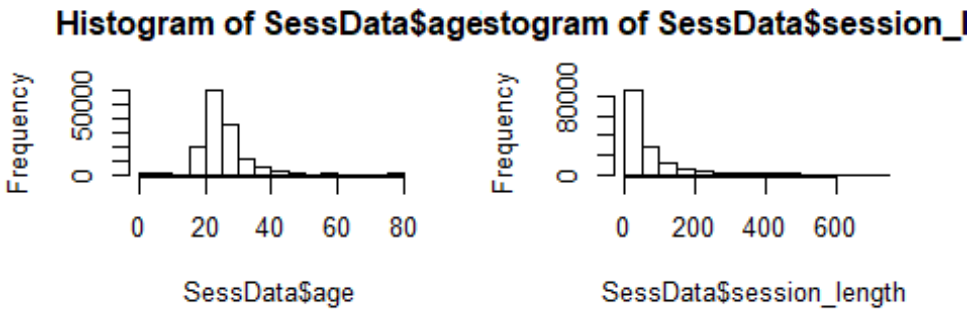
```r
hist(SessData$session_length)

train_data <- sqldf("select * from SessData where startdate < '2009-04-01'")
test_data <-  sqldf("select * from SessData where startdate > '2009-03-31'")
panel.data.train <- plm.data(train_data, index = c("session_start","userid"))

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

panel.data.test <- plm.data(test_data, index = c("session_start","userid"))

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

str(panel.data.train)

## Classes 'plm.dim' and 'data.frame':  129732 obs. of  10 variables:
##  $ session_start : Factor w/ 129633 levels "2005-02-14 00:02:10",..: 1 2 3
4 5 6 7 8 9 10 ...
##  $ userid        : Factor w/ 158 levels "user_000009",..: 124 124 124 124
124 76 124 124 124 69 ...
##  $ gender        : Factor w/ 3 levels "","f","m": 2 2 2 2 2 3 2 2 2 3 ...
##  $ age           : int  23 23 23 23 23 32 23 23 23 33 ...
##  $ country       : Factor w/ 56 levels "","Antarctica",..: 53 53 53 53 53
52 53 53 53 1 ...
##  $ startdate     : Factor w/ 1586 levels "2005-02-14","2005-02-15",..: 1 1
1 1 2 2 2 3 3 3 ...
##  $ day_of_week   : Factor w/ 7 levels "friday   ","monday   ",..: 2 2 2 2
6 6 6 7 7 7 ...
##  $ timeofday     : Factor w/ 4 levels "evening","morning",..: 3 3 2 1 1 1
1 3 3 4 ...
##  $ sessionid     : int  1 2 3 4 1 1 2 1 2 1 ...
##  $ session_length: int  3 108 24 158 23 10 38 35 116 8 ...

summary(panel.data.train)

##               session_start            userid       gender
##  2006-04-28 22:55:16:     2   user_000089:  4654    : 2851
##  2006-05-03 19:51:47:     2   user_000084:  3370   f:53657
##  2006-05-31 08:20:23:     2   user_000188:  3280   m:73224
##  2006-06-04 21:46:30:     2   user_000215:  3142
##  2006-06-21 00:24:39:     2   user_000242:  2691
##  2006-06-24 11:04:18:     2   user_000296:  2542
##  (Other)            :129720   (Other)    :110053
##       age                  country           startdate
##  Min.   : 3.00   United States :17095   2009-02-05:   175
##  1st Qu.:22.00   United Kingdom:13992   2009-03-17:   175
##  Median :24.00   Poland        :13305   2009-02-26:   173
##  Mean   :25.62   Turkey        : 9545   2009-03-13:   173
##  3rd Qu.:28.00   Italy         : 9348   2008-11-25:   171
##  Max.   :77.00   Sweden        : 6866   2009-02-09:   168
```

```
##                   (Other)       :59581   (Other)   :128697
##     day_of_week      timeofday       sessionid      session_length
##  friday   :17893   evening:36636   Min.   : 1.000   Min.   :  3.00
##  monday   :19446   morning:32213   1st Qu.: 1.000   1st Qu.: 10.00
##  saturday :17222   night  :23956   Median : 2.000   Median : 32.00
##  sunday   :18155   noon   :36927   Mean   : 2.098   Mean   : 52.33
##  thursday :19011                   3rd Qu.: 3.000   3rd Qu.: 71.00
##  tuesday  :19160                   Max.   :14.000   Max.   :720.00
##  wednesday:18845
```

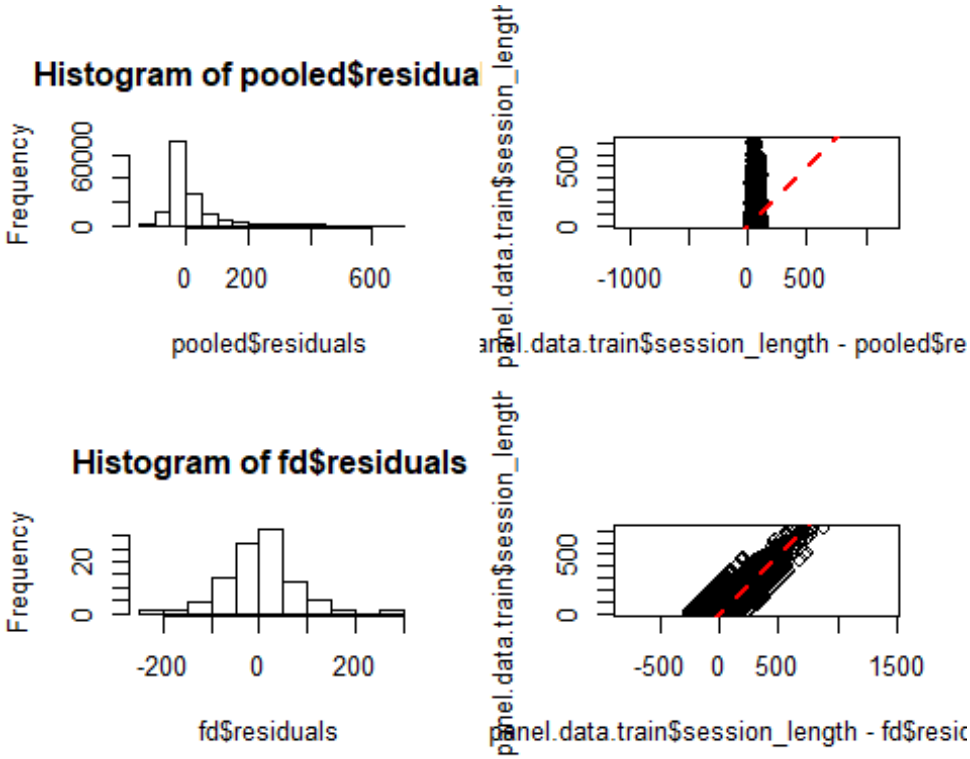
```
pooled <-plm(session_length~gender+age+timeofday+day_of_week+country,  data =
panel.data.train, model = "pooling")

fd <-plm(session_length~gender+age+timeofday+day_of_week+country, data =
panel.data.train, model = "fd")

fe <-plm(session_length~gender+age+timeofday+day_of_week+country, data =
panel.data.train, model = "within")

random <-plm(session_length~gender+age+timeofday+day_of_week+country, data =
panel.data.train, model = "random")

##To decide between the fixed and random effects model we will run the
Hausman test

phtest(fe, random)
```

```
##
##  Hausman Test
##
## data:  session_length ~ gender + age + timeofday + day_of_week + country
## chisq = 40.308, df = 34, p-value = 0.2113
## alternative hypothesis: one model is inconsistent
```

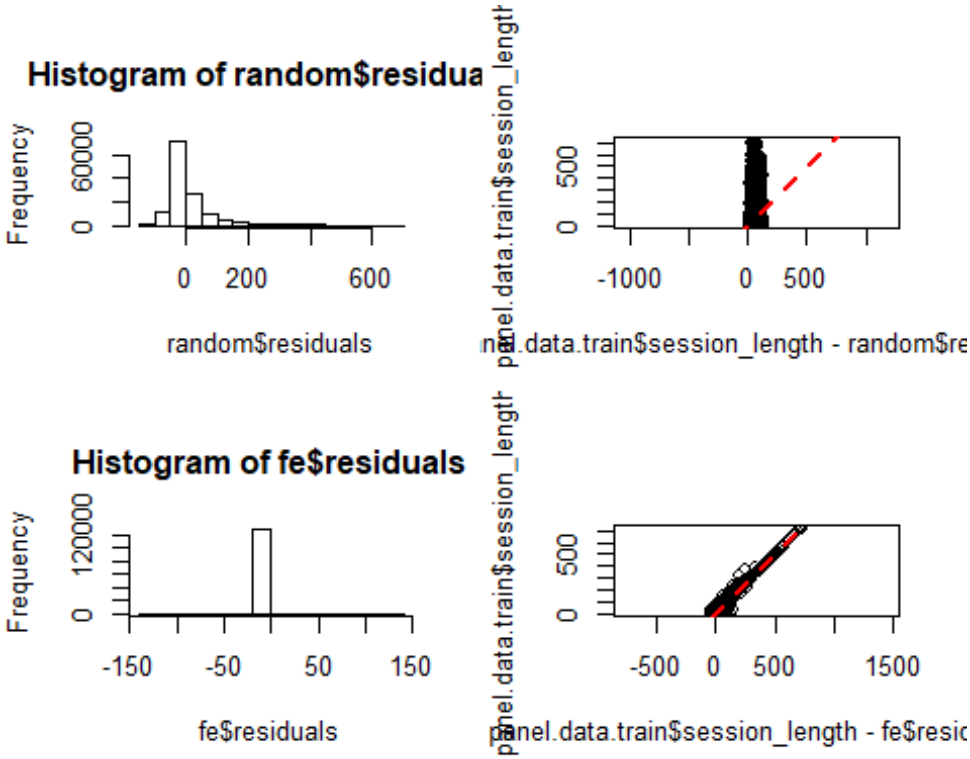
```
par(mfrow=c(2, 2))
```




```
hist(pooled$residuals)
# Fitted-vs-observed plot
plot(panel.data.train$session_length - pooled$residuals,
panel.data.train$session_length, asp = 1)
abline(0, 1, col = 'red', lty = 'dashed', lwd = 2)

hist(fd$residuals)
# Fitted-vs-observed plot
plot(panel.data.train$session_length - fd$residuals,
panel.data.train$session_length, asp = 1)

## Warning in panel.data.train$session_length - fd$residuals: longer object
## length is not a multiple of shorter object length

abline(0, 1, col = 'red', lty = 'dashed', lwd = 2)
```

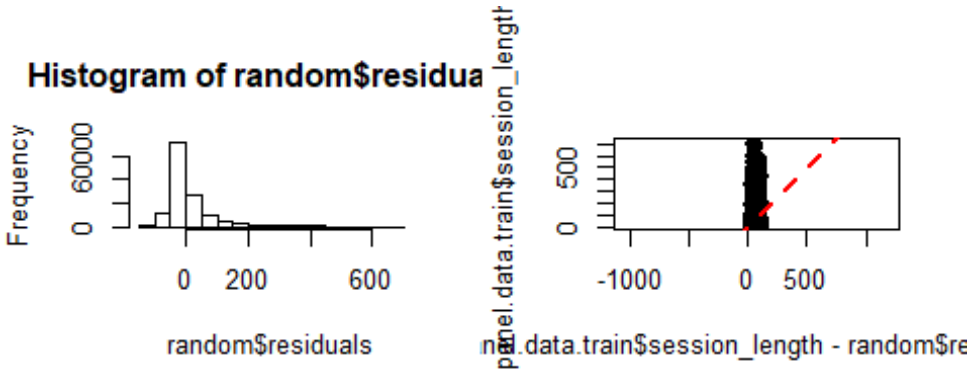



```r
par(mfrow=c(2, 2))

hist(random$residuals)
# Fitted-vs-observed plot
plot(panel.data.train$session_length - random$residuals,
panel.data.train$session_length, asp = 1)
abline(0, 1, col = 'red', lty = 'dashed', lwd = 2)

hist(fe$residuals)
# Fitted-vs-observed plot
plot(panel.data.train$session_length - fe$residuals,
panel.data.train$session_length, asp = 1)
abline(0, 1, col = 'red', lty = 'dashed', lwd = 2)
```




```
hist(random$residuals)
# Fitted-vs-observed plot
plot(panel.data.train$session_length - random$residuals,
panel.data.train$session_length, asp = 1)
abline(0, 1, col = 'red', lty = 'dashed', lwd = 2)

# Mean absolute error
mae(panel.data.train$session_length - pooled$residuals,
panel.data.train$session_length)

## [1] 42.47056

mae(panel.data.train$session_length - fd$residuals,
panel.data.train$session_length)

## Warning in panel.data.train$session_length - fd$residuals: longer object
## length is not a multiple of shorter object length

## [1] 50.55785

mae(panel.data.train$session_length - random$residuals,
panel.data.train$session_length)

## [1] 42.47056

mae(panel.data.train$session_length - fe$residuals,
panel.data.train$session_length)
```

```
## [1] 0.03857971
```


Histogram of random$residua
Frequency
0
60000
0
200
600
random$residuals
panel.data.train$session_length
0
500
-1000
0
500
nel.data.train$session_length - random$re