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(utils)
library(stringr)
```

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

```
library(prediction)
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
```

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
train_data <- na.omit(read.csv(file="usersessions-with-genre-train.csv", header=TRUE, row.names = NULL,

test_data <- na.omit(read.csv(file="usersessions-with-genre-test.csv", header=TRUE, row.names = NULL, s

train_data$id <- as.numeric(str_replace_all(train_data$userid, "user_", ""))
test_data$id <- as.numeric(str_replace_all(test_data$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
```

```r
nrow(train_data)
```

```
## [1] 169556
```

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

##Summaries

summary(mdl_fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = session_length ~ age + session_length_mvavg + previous_duration +
##     is_holiday + absence_time, data = panel.data.train, model = "within")
##
## Unbalanced Panel: n = 174, T = 19-4312, N = 169556
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -99332.1  -3343.7  -1450.5   1498.3 567439.3
##
## Coefficients:
##                          Estimate  Std. Error t-value  Pr(>|t|)
## age                   -1.5402e+02  2.4207e+01 -6.3626 1.989e-10 ***
## session_length_mvavg   6.5481e-01  1.6699e-02 39.2122 < 2.2e-16 ***
## previous_duration      1.4451e-01  2.3814e-03 60.6825 < 2.2e-16 ***
## is_holiday            -3.8898e+00  5.8847e+01 -0.0661    0.9473
## absence_time          -9.0199e-06  4.6441e-05 -0.1942    0.8460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1.3296e+13
## Residual Sum of Squares: 1.2818e+13
## R-Squared:      0.03597
## Adj. R-Squared: 0.034957
## F-statistic: 1263.97 on 5 and 169377 DF, p-value: < 2.22e-16
```

```r
#cat(length(panel.data.train$session_length), length(mdl_fd$residuals))
# Fitted vs Observed and Fitted vs Residuals plots
```

```r
# par(mfrow=c(1,2))
# plot(mdl_fe$fitted.values, mdl_fe$model$session_length, asp=1, ylab = "Observed", xlab = "Fitted", ma
# abline(0,1, col='red', lty='dashed', lwd=2)
#
# ## Fitted vs Residuals plots
# plot(mdl_fe$fitted.values,mdl_fe$residuals, asp=1, ylab = "Residuals", xlab = "Fitted", main = "Fitte
# abline(0,0, col='red', lty='dashed', lwd=2)
#
# fitval <- prediction(mdl_fe, data = find_data(mdl_fe, parent.frame()), at=NULL, calculate_se=FALSE)


# ## Fitted vs Fitted by hand plots
intercept <- within_intercept(mdl_fe)
fixefs <- as.numeric(fixef(mdl_fe)[index(mdl_fe, which = "id")])

cat('Overall intercept ', intercept)
```

```
## Overall intercept  5063.798
```

```r
# cat('Fixed effects ', fixefs)
cat('Coeff - Age ', mdl_fe$coefficients["age"])
```

```
## Coeff - Age  -154.0199
```

```r
cat('Coeff - Moving average ', mdl_fe$coefficients["session_length_mvavg"])
```

```
## Coeff - Moving average  0.6548119
```

```r
cat('Coeff - Previous duration ', mdl_fe$coefficients["previous_duration"])
```

```
## Coeff - Previous duration  0.144511
```

```r
temp <- merge(panel.data.train, data.frame(fixef_usr = names(fixef(mdl_fe)), fixef = as.numeric(fixef(m

fit_hand <- temp$fixef + as.numeric(mdl_fe$coefficients[1])* panel.data.train$age +  as.numeric(mdl_fe$
 as.numeric(mdl_fe$coefficients[3]) * panel.data.train$previous_duration +
 as.numeric(mdl_fe$coefficients[4]) * panel.data.train$is_holiday +
 as.numeric(mdl_fe$coefficients[5]) * panel.data.train$absence_time

head(fit_hand)
```

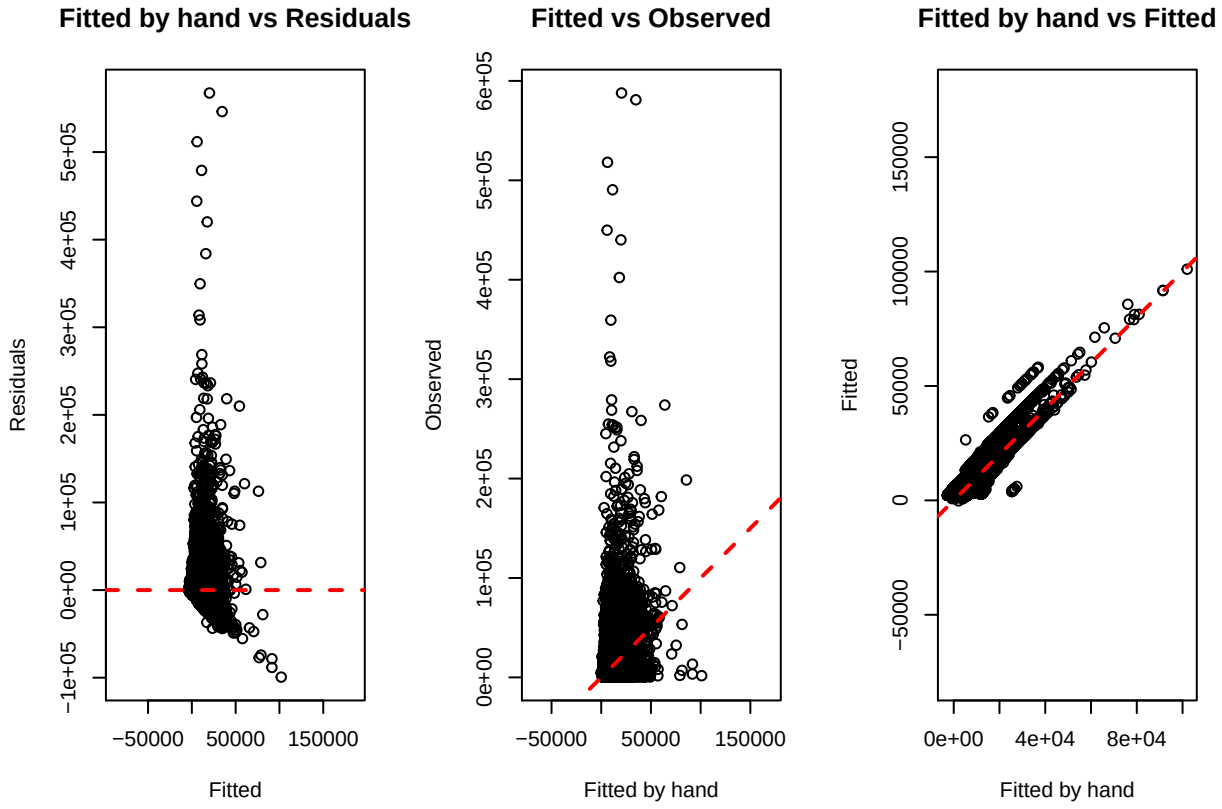
```
## [1] 6382.533 5828.086 8314.609 5546.719 5058.564 4567.493
```

```r
par(mfrow=c(1,3))
plot(fit_hand, mdl_fe$residuals,asp=1, ylab = "Residuals", xlab = "Fitted", main = "Fitted by hand vs R
abline(0,0, col='red', lty='dashed', lwd=2)

# # ## Observed vs Fitted
plot(panel.data.train$session_length-mdl_fe$residuals,panel.data.train$session_length, asp=1, xlab = "F
```

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

# # ## Fitted by hand vs Fitted plots
plot(fit_hand,panel.data.train$session_length-mdl_fe$residuals, asp=1, xlab = "Fitted by hand", ylab =
abline(0,1, col='red', lty='dashed', lwd=2)
```




```r
# ## MAE and RMSE
#
# mae_fe = mean(abs(mdl_fe$residuals))
# rmse_fe = sqrt(mean(abs(mdl_fe$residuals)^2))
#
mae_fe = mean(abs(panel.data.train$session_length-fit_hand))
rmse_fe = sqrt(mean(abs(panel.data.train$session_length-fit_hand)^2))

cat('MAE = ', mae_fe, ', RMSE = ', rmse_fe)
```

```
## MAE =  4584.935 , RMSE =  8924.405
```

```r
## Out-of-Sample Predictions on test data
#fxeff <- data.frame(fixef_usr= names(fixef(mdl_fe)), fixef = as.numeric(fixef(mdl_fe)[index(mdl_fe, wh
#temp <- merge(test_data, data.frame(fixef_usr= names(fixef(mdl_fe)), fixef = as.numeric(fixef(mdl_fe))

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

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

```r
temp <- merge(panel.data.test, data.frame(fixef_usr = names(fixef(mdl_fe)), fixef = as.numeric(fixef(mdl

predval <- temp$fixef + as.numeric(mdl_fe$coefficients[1])* panel.data.test$age +  as.numeric(mdl_fe$coe
 as.numeric(mdl_fe$coefficients[3]) * panel.data.test$previous_duration +
 as.numeric(mdl_fe$coefficients[4]) * panel.data.test$is_holiday +
 as.numeric(mdl_fe$coefficients[5]) * panel.data.test$absence_time

pred<-cbind(predval, panel.data.test$session_length)
# # ## Observed vs Predicted plots
plot(predval,panel.data.test$session_length, asp=1, xlab = "Predicted", ylab = "Observed", main = "Predi
abline(0,1, col='red', lty='dashed', lwd=2)
# ## MAE and RMSE
#
mae_fe = mean(abs(mdl_fe$residuals))
rmse_fe = sqrt(mean(abs(mdl_fe$residuals)^2))
cat('Fitted - MAE = ', mae_fe, ', RMSE = ', rmse_fe)
```

```
## Fitted - MAE =  4311.5 , RMSE =  8694.647
```

```r
#
mae_pred = mean(abs(panel.data.test$session_length-predval))
rmse_pred = sqrt(mean(abs(panel.data.test$session_length-predval)^2))
cat('Predicted - MAE = ', mae_pred, ', RMSE = ', rmse_pred)
```

```
## Predicted - MAE =  4153.83 , RMSE =  7830.161
```