Panel Data - Random Effects

Jayashree Raman

12/11/2018

```
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(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(stringr)
```

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

```
library(ggplot2)
```

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

```
library(dplyr)
```

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plm':
##
##     between, lag, lead
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gridExtra)
```

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

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

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:

Cluster the users by average session_length and absence_time

```
cluster_kmeans <- function(sessdf){

  sessdf %>% mutate_if(is.numeric, scale)
  clustdat <- sessdf[,2:4]
  k5 <- kmeans(clustdat, centers = 5, nstart = 25)
  clustdat <- cbind(sessdf, cluster = k5$cluster)
  return(clustdat)
}
```

Include cluster number in dataset

```
SessData <- na.omit(read.csv(file="usersessions-avgbase.csv", header=TRUE, row.names = NULL, sep="|")[,
SessData$session_dt <- as.Date(SessData$session_dt,format= "%Y-%m-%d")
SessData$id <- as.numeric(str_replace_all(SessData$userid, 'user_', ''))
SessData$previous_duration_scale <- scale(SessData$previous_duration)
SessData$absence_time_scale <- scale(SessData$absence_time)
SessData$session_mvavg_scale <- scale(SessData$session_length_mvavg)

SessData$previous_duration_log <- log(SessData$previous_duration)
SessData$absence_time_log <- log(SessData$absence_time)
SessData$session_mvavg_log <- log(SessData$session_length_mvavg)
SessData$avg_base_log <- log(SessData$avg_base)
SessData$session_length_log <- log(SessData$session_length)
```

```
user_cluster <- cluster_kmeans(sqldf("select userid, avg(session_length) sess_length_avg, avg(absence_t
head(user_cluster)
```

```
##        userid sess_length_avg absence_time_avg sess_count cluster
## 1 user_000001        4933.335         76791.70        984       1
## 2 user_000002        6614.383         44002.40       1820       1
## 3 user_000003        4300.593         74145.36       1157       1
## 4 user_000005        4891.373         39141.03       1237       1
## 5 user_000006        5077.173         34326.90       2010       1
## 6 user_000008       12287.429         25805.98        249       1
```

```
user_cluster$id <- as.numeric(str_replace_all(user_cluster$userid, 'user_', ''))

session_cluster <- sqldf("select a.*, b.cluster from SessData a, user_cluster b where a.userid = b.user

head(session_cluster)
```

```
##        userid       session_start session_dt session_length absence_time
## 1 user_000001 2006-08-15 12:27:51 2006-08-15          11200        67965
## 2 user_000001 2006-08-16 13:43:02 2006-08-16           5778        80047
## 3 user_000001 2006-08-18 19:09:27 2006-08-18           3080        59165
## 4 user_000001 2006-08-19 12:26:52 2006-08-19           7955        80547
## 5 user_000001 2006-08-21 03:36:08 2006-08-21           1160         2272
## 6 user_000001 2006-08-21 04:33:20 2006-08-21           1375         2190
##   previous_duration previous_absence_time session_length_mvavg
## 1             15305                152007             15305.00
## 2              2958                  8788              9821.00
## 3              8125                 98435              8673.20
## 4              3080                 59165              7741.00
## 5              7726                 44728              7765.88
## 6              1160                  2272              7031.89
##   session_length_mvavg_30 session_length_mvavg_time sessionid avg_base id
## 1                15305.00                  15305.00         2 4827.997  1
## 2                 9821.00                   2958.00         4 4827.997  1
## 3                 8673.20                  13252.50         6 4827.997  1
## 4                 7741.00                   4368.00         7 4827.997  1
## 5                 7765.88                   7726.00         9 4827.997  1
## 6                 7031.89                   9861.67        10 4827.997  1
##   previous_duration_scale absence_time_scale session_mvavg_scale
## 1               1.1279262         0.03391953          1.99706413
## 2              -0.4113398         0.06479825          0.73222227
## 3               0.2328156         0.01142883          0.46749115
## 4              -0.3961304         0.06607613          0.25248648
## 5               0.1830734        -0.13397614          0.25822486
## 6              -0.6354915        -0.13418571          0.08893579
##   previous_duration_log absence_time_log session_mvavg_log avg_base_log
## 1              9.635935        11.126748          9.635935     8.482187
## 2              7.992269        11.290369          9.192278     8.482187
## 3              9.002701        10.988085          9.067993     8.482187
## 4              8.032685        11.296596          8.954286     8.482187
## 5              8.952347         7.728416          8.957495     8.482187
## 6              7.056175         7.691657          8.858211     8.482187
```

```
##   session_length_log cluster
## 1           9.323669       1
## 2           8.661813       1
## 3           8.032685       1
## 4           8.981556       1
## 5           7.056175       1
## 6           7.226209       1
```

```
cluster_count <- sqldf("select cluster, count(distinct userid) no_users from session_cluster group by cl
nrow(session_cluster)
```

```
## [1] 633823
```

```
rm(SessData)

results<-setNames(data.frame(matrix(ncol = 5, nrow = 0)), c("model","cluster", "session_length", "sessio
mae_results<-setNames(data.frame(matrix(ncol = 16, nrow = 0)), c("cluster", "no_users","base", "base_log
results <- data.frame()
results_scale <- data.frame()
```

```
for (i in 0:nrow(cluster_count))
{
if( i > 0 ){
     df_train <- subset(session_cluster, cluster == i & session_dt < "2009-04-01")
     df_test <- subset(session_cluster, cluster == i & session_dt >= "2009-04-01")
     no_users <- as.numeric(cluster_count[i, "no_users"])
     mae_baseline <-  mean(abs(df_test[df_test$cluster==i,]$session_length-df_test[df_test$cluster==i,]$
     mae_baseline_log <-  mean(abs(df_test[df_test$cluster==i,]$session_length_log-df_test[df_test$clust
  }
  else
  {
    df_train <- subset(session_cluster, session_dt < "2009-04-01")
    df_test <- subset(session_cluster, session_dt >= "2009-04-01")
    no_users <- nrow(user_cluster)
    mae_baseline <-  mean(abs(df_test$session_length-df_test$avg_base))
    mae_baseline_log <-  mean(abs(df_test$session_length_log-df_test$avg_base_log))
    }

  mae_pred <- 'NA'
  rmse_pred <- 'NA'

  if(nrow(df_train)>0 & nrow(df_test)>0 & no_users>1)
    {
    mdl_random <-plm(session_length~ previous_duration+absence_time+session_length_mvavg+session_length_
    mdl_random_scale <-plm(session_length~ previous_duration_scale + absence_time_scale+session_mvavg_sc
    mdl_random_log <-plm(session_length_log~previous_duration_log + absence_time_log+session_mvavg_log,
   predval <- prediction(mdl_random, data=df_test)
   results <- rbind(results, predval)
   predval_scale <- prediction(mdl_random_scale, data=df_test)
   predval_log <- prediction(mdl_random_log, data=df_test)
 if(i >0)
    {
      results_scale <- rbind(results_scale, predval_scale)
```

```
      results <- rbind(results, predval)
   }

   mae_pred <- mean(abs(predval$session_length-predval$fitted))
   rmse_pred<-sqrt(mean(abs(predval$session_length-predval$fitted)^2))
   mae_pred_scale<-mean(abs(predval_scale$session_length-predval_scale$fitted))
   rmse_pred_scale<-sqrt(mean(abs(predval_scale$session_length-predval_scale$fitted)^2))
   mae_pred_log<-mean(abs(predval_log$session_length_log-predval_log$fitted))
   rmse_pred_log<-sqrt(mean(abs(predval_log$session_length_log-predval_log$fitted)^2))
  }

 if(is.numeric(mae_pred)){
mae_results[nrow(mae_results)+1,]<-c(i, no_users, mae_baseline, mae_baseline_log,  mae_pred, rmse_pred,
mae_pred_log/mae_baseline_log, (mae_pred_log/mae_baseline_log)*no_users
  }
  else
  {
    mae_results[nrow(mae_results)+1,]<-c(i, no_users,mae_baseline, mae_baseline_log, 0, 0, 0, 0, 0, 0, 0
  }

rm(df_train)
rm(df_test)
}

results$residuals <- results$session_length - results$fitted
results_scale$residuals <- results_scale$session_length - results_scale$fitted

sqldf("select cluster id, no_users users, round(base, 4) base, round(base_log, 4) base_log, round(mae, 4
round(norm_mae_log_tot, 4) nrm_log_tot from mae_results")
```

```
##   id users     base base_log      mae mae_scale mae_log nrm_mae
## 1  0   620 4644.348   0.9271 4292.093  4435.338  0.8265  0.9242
## 2  1   446 4687.439   0.9268 4353.306  4495.925  0.8279  0.9287
## 3  2     1 2146.161   1.2751    0.000     0.000  0.0000  0.0000
## 4  3    33 4868.590   0.9382 3536.704  3737.295  0.7697  0.7264
## 5  4    12 2808.919   0.8317 2828.605  2739.785  0.8380  1.0070
## 6  5   128 4242.937   0.9298 3840.547  3950.345  0.8184  0.9052
##   nrm_mae_scl nrm_mae_log  nrm_tot nrm_scale_tot nrm_log_tot
## 1      0.9550      0.8915 572.9756      592.0981    552.7178
## 2      0.9591      0.8933 414.2080      427.7779    398.3914
## 3      0.0000      0.0000   0.0000        0.0000      0.0000
## 4      0.7676      0.8204  23.9723       25.3319     27.0746
## 5      0.9754      1.0075  12.0841       11.7047     12.0904
## 6      0.9310      0.8801 115.8608      119.1731    112.6556
```

```
mae_full <- mae_results[mae_results$cluster == 0,]$norm_mae_tot/mae_results[mae_results$cluster == 0,]$r
mae_scale_full <- mae_results[mae_results$cluster == 0,]$norm_mae_scale_tot/mae_results[mae_results$clus
mae_log_full <- mae_results[mae_results$cluster == 0,]$norm_mae_log_tot/mae_results[mae_results$cluster

mae_cluster <- sum(mae_results[mae_results$cluster > 0,]$norm_mae_tot, na.rm = TRUE)/sum(mae_results[mae
mae_cluster_scale <- sum(mae_results[mae_results$cluster > 0,]$norm_mae_scale_tot, na.rm =TRUE)/sum(mae_
mae_cluster_log <- sum(mae_results[mae_results$cluster > 0,]$norm_mae_log_tot, na.rm =TRUE)/sum(mae_resu
```

```
cat('Full MAE ', mae_full, ' Scaled MAE ', mae_scale_full, ' Log MAE ', mae_log_full)
```

```
## Full MAE  0.9241542  Scaled MAE  0.9549969  Log MAE  0.8914802
```

```
cat(' Clustered MAE ', mae_cluster,' Clustered Scaled MAE ', mae_cluster_scale, ' Clustered Log MAE ', m
```

```
##  Clustered MAE  0.9131051  Clustered Scaled MAE  0.9419155  Clustered Log MAE  0.8874386
```

Plot clusters

```
head(results)
```

```
##        userid       session_start session_dt session_length absence_time
## 1 user_000001 2009-04-01 01:19:04 2009-04-01           1102        38821
## 2 user_000001 2009-04-01 12:24:27 2009-04-01          13314        37114
## 3 user_000001 2009-04-02 02:24:55 2009-04-02           1599         7238
## 4 user_000001 2009-04-02 11:33:40 2009-04-02           3243         3029
## 5 user_000001 2009-04-02 13:18:12 2009-04-02           2195         2992
## 6 user_000001 2009-04-02 14:44:39 2009-04-02           5462        59278
##   previous_duration previous_absence_time session_length_mvavg
## 1              9317                 20806             4828.00
## 2              1102                 38821             4823.94
## 3             13314                 37114             4833.18
## 4              1599                  7238             4829.66
## 5              3243                  3029             4827.94
## 6              2195                  2992             4825.08
##   session_length_mvavg_30 session_length_mvavg_time sessionid avg_base id
## 1                 5411.20                   8880.0       918 4827.997  1
## 2                 5393.20                   5201.8       919 4827.997  1
## 3                 5569.90                   7708.6       920 4827.997  1
## 4                 5477.53                   4881.6       921 4827.997  1
## 5                 5537.17                   4697.6       922 4827.997  1
## 6                 5518.07                   7354.0       923 4827.997  1
##   previous_duration_scale absence_time_scale session_mvavg_scale
## 1               0.3814190       -0.04056559          -0.4193743
## 2              -0.6427222       -0.04492828          -0.4203107
## 3               0.8797138       -0.12128422          -0.4181795
## 4              -0.5807626       -0.13204143          -0.4189914
## 5              -0.3758097       -0.13213599          -0.4193881
## 6              -0.5064609        0.01171763          -0.4200477
##   previous_duration_log absence_time_log session_mvavg_log avg_base_log
## 1              9.139596        10.566717          8.482188     8.482187
## 2              7.004882        10.521750          8.481346     8.482187
## 3              9.496571         8.887100          8.483260     8.482187
## 4              7.377134         8.015988          8.482531     8.482187
## 5              8.084254         8.003697          8.482175     8.482187
## 6              7.693937        10.989994          8.481583     8.482187
##   session_length_log cluster   fitted se.fitted   residuals
## 1           7.004882       1 6962.933        NA -5860.93267
## 2           9.496571       1 4533.150        NA  8780.84982
```

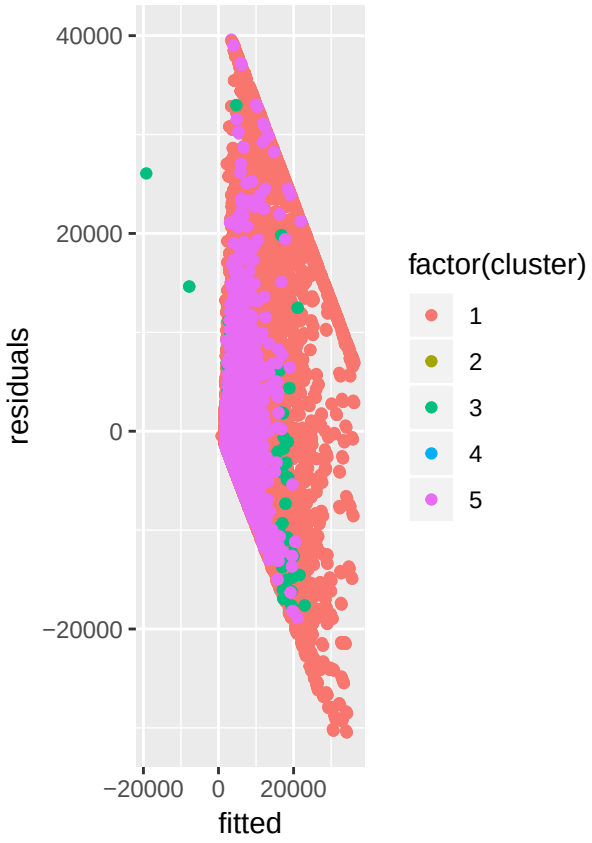
```
## 3           7.377134       1 6926.056        NA -5327.05560
## 4           8.084254       1 4463.353        NA -1220.35323
## 5           7.693937       1 4567.790        NA -2372.78996
## 6           8.605570       1 5541.342        NA   -79.34202
```

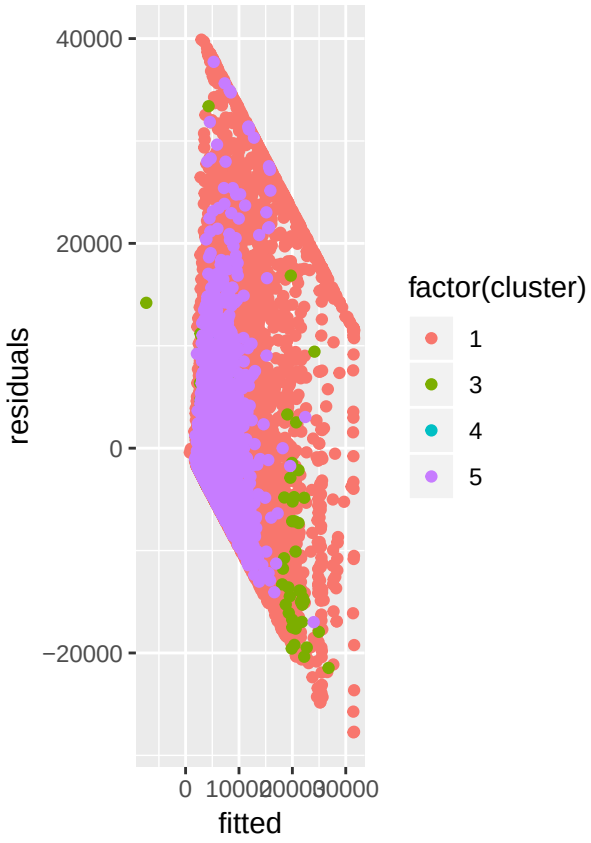
```
head(results_scale)
```

```
##        userid       session_start session_dt session_length absence_time
## 1 user_000001 2009-04-01 01:19:04 2009-04-01           1102        38821
## 2 user_000001 2009-04-01 12:24:27 2009-04-01          13314        37114
## 3 user_000001 2009-04-02 02:24:55 2009-04-02           1599         7238
## 4 user_000001 2009-04-02 11:33:40 2009-04-02           3243         3029
## 5 user_000001 2009-04-02 13:18:12 2009-04-02           2195         2992
## 6 user_000001 2009-04-02 14:44:39 2009-04-02           5462        59278
##   previous_duration previous_absence_time session_length_mvavg
## 1              9317                 20806             4828.00
## 2              1102                 38821             4823.94
## 3             13314                 37114             4833.18
## 4              1599                  7238             4829.66
## 5              3243                  3029             4827.94
## 6              2195                  2992             4825.08
##   session_length_mvavg_30 session_length_mvavg_time sessionid avg_base id
## 1                5411.20                   8880.0       918 4827.997  1
## 2                5393.20                   5201.8       919 4827.997  1
## 3                5569.90                   7708.6       920 4827.997  1
## 4                5477.53                   4881.6       921 4827.997  1
## 5                5537.17                   4697.6       922 4827.997  1
## 6                5518.07                   7354.0       923 4827.997  1
##   previous_duration_scale absence_time_scale session_mvavg_scale
## 1               0.3814190        -0.04056559          -0.4193743
## 2              -0.6427222        -0.04492828          -0.4203107
## 3               0.8797138        -0.12128422          -0.4181795
## 4              -0.5807626        -0.13204143          -0.4189914
## 5              -0.3758097        -0.13213599          -0.4193881
## 6              -0.5064609         0.01171763          -0.4200477
##   previous_duration_log absence_time_log session_mvavg_log avg_base_log
## 1              9.139596        10.566717          8.482188     8.482187
## 2              7.004882        10.521750          8.481346     8.482187
## 3              9.496571         8.887100          8.483260     8.482187
## 4              7.377134         8.015988          8.482531     8.482187
## 5              8.084254         8.003697          8.482175     8.482187
## 6              7.693937        10.989994          8.481583     8.482187
##   session_length_log cluster   fitted se.fitted  residuals
## 1           7.004882       1 5495.452        NA -4393.4516
## 2           9.496571       1 4095.548        NA  9218.4523
## 3           7.377134       1 6195.228        NA -4596.2278
## 4           8.084254       1 4201.408        NA  -958.4082
## 5           7.693937       1 4479.947        NA -2284.9467
## 6           8.605570       1 4271.151        NA  1190.8490
```

```
residplot <- ggplot(results) + geom_point(aes(x = fitted, y = residuals, colour = factor(cluster)))

residplot_scale <- ggplot(results_scale) + geom_point(aes(x = fitted, y = residuals, colour = factor(clu
grid.arrange(residplot, residplot_scale, ncol=2)
```


40000
20000
0
−20000
residuals
−20000
0
20000
fitted
factor(cluster)
1
2
3
4
5


40000
20000
0
−20000
residuals
0
10000
20000
30000
fitted
factor(cluster)
1
3
4
5