Lab work 2

Artem Storozhuk

IMPORTING DATA

First of all, lets read our data and transform it for further use:

```
filenames = list.files(path='data', full.names=TRUE)
datalist = lapply(filenames,function(x){
  x0 <- read.csv(file = x,header = F)[,c(1,6)];
  colnames(x0) <- c("data", unlist(strsplit(x,"[_.]"))[2]);x0})
y <- Reduce(function(x,y){
  merge(x,y,by="data")
  },datalist)
Data <- y[-nrow(y),-1]
Data$adi <- y$adi[-1]
nn <- nrow(Data)
```

FIRST MODEL

Consider the model using all data except the last 20 sessions:

```
model1<-lm(adi~.-adi, data = Data[1:(nn-20),])
summary(model1)
```

```
##
## Call:
## lm(formula = adi ~ . - adi, data = Data[1:(nn - 20), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0323 -0.5631 -0.1186  0.5082  1.7388
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.67982    3.49274   2.485 0.013844 *
## adm          0.49010    0.08046   6.092 6.39e-09 ***
## adp          0.11484    0.07437   1.544 0.124289
## adsk         0.31152    0.05481   5.684 5.08e-08 ***
## adt          0.18530    0.04094   4.526 1.08e-05 ***
## aee         -0.18579    0.11615  -1.600 0.111408
## aep          0.02186    0.11112   0.197 0.844248
## aes         -1.05038    0.29117  -3.607 0.000398 ***
## aet          0.13571    0.03623   3.745 0.000241 ***
## afl          0.06326    0.04782   1.323 0.187521
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.769 on 184 degrees of freedom
## Multiple R-squared:  0.9272, Adjusted R-squared:  0.9236
## F-statistic: 260.4 on 9 and 184 DF,  p-value: < 2.2e-16
```

We got very good result, determination coefficient equals 0.9236. Also, p-value < 2.2e-16, which means that there is a dependency between feedback and regressors.

SECOND MODEL

Next, consider the model using 50 sessions before last 20 sessions:

```
model2<-lm(adi~adm+adsk+adt+aes+aet, data = Data[(nn-70):(nn-20),])
summary(model2)
```

```
##
## Call:
## lm(formula = adi ~ adm + adsk + adt + aes + aet, data = Data[(nn -
##     70):(nn - 20), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60943 -0.41246 -0.06408  0.39117  1.41818
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.06187    8.08254   2.235  0.03045 *
## adm          0.60520    0.17953   3.371  0.00155 **
## adsk         0.31724    0.13323   2.381  0.02155 *
## adt          0.06930    0.13566   0.511  0.61196
## aes         -0.93889    0.35564  -2.640  0.01135 *
## aet          0.07248    0.11480   0.631  0.53099
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6975 on 45 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.3812
## F-statistic: 7.162 on 5 and 45 DF,  p-value: 5.29e-05
```

We can see that R-squared is much worse than in the previous model, but p-value = 5.29e-05 indicates that there is a dependency between feedback and regressors.

PCA

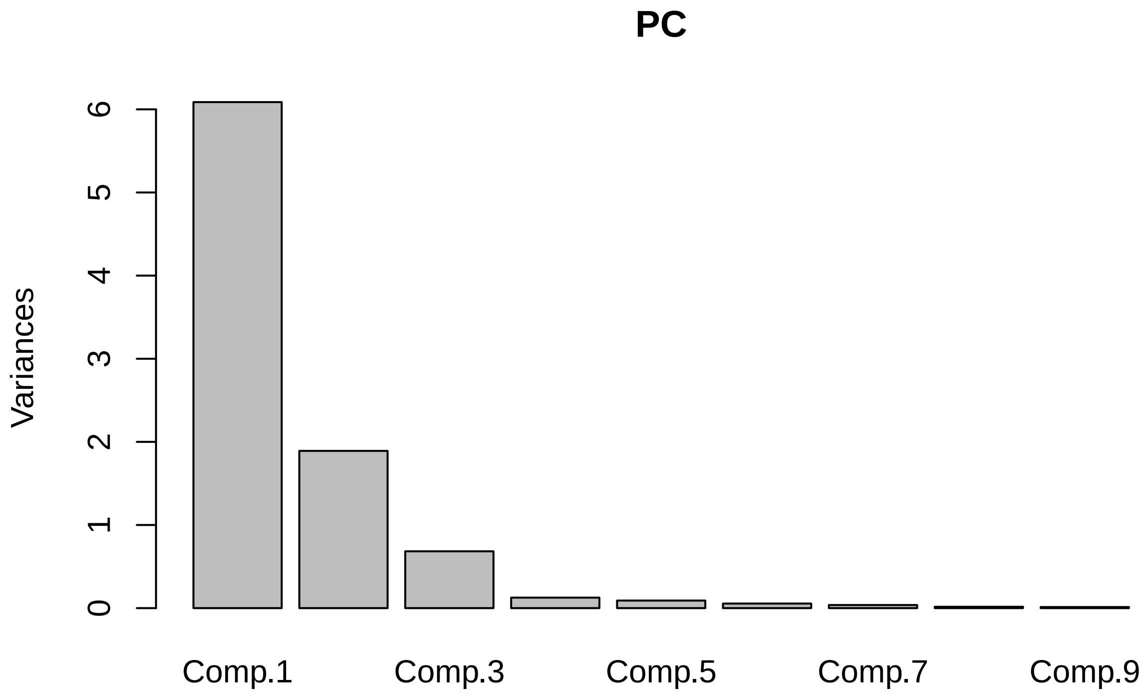
Lets carry out Principal component analysis. For this, we need to create them:

```
X <- Data[-1]
PC <- princomp(X, cor=T)
summary(PC)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     2.4671849 1.3752522 0.82679921 0.35503376 0.30137848
## Proportion of Variance 0.6763335 0.2101465 0.07595522 0.01400544 0.01009211
## Cumulative Proportion  0.6763335 0.8864800 0.96243521 0.97644065 0.98653276
##                             Comp.6      Comp.7      Comp.8      Comp.9
## Standard deviation     0.234322311 0.193295599 0.126259220 0.113989631
## Proportion of Variance 0.006100772 0.004151465 0.001771266 0.001443737
## Cumulative Proportion  0.992633532 0.996784997 0.998556263 1.000000000
```

Eigenvalues diagram:

```
plot(PC)
```



Consider the load on the main components:

```
loadings(PC)
```

```
##
## Loadings:
##      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## adm   0.393                0.463  0.317  0.233  0.564  0.303  0.246
## adp   0.390  0.166         0.206        -0.222  0.126 -0.215 -0.812
## adsk  0.239 -0.554  0.155  0.149 -0.745  0.161
## adt         -0.658  0.475         0.529 -0.163 -0.121 -0.125
## aee   0.387        -0.267  0.318  0.160  0.382 -0.634 -0.302  0.112
## aep   0.351 -0.229 -0.420 -0.392  0.145        -0.158  0.635 -0.191
## aes   0.384 -0.111 -0.213 -0.529                0.385 -0.565  0.211
## aet   0.377  0.203  0.181        -0.116 -0.731 -0.215         0.424
## afl   0.272  0.352  0.645 -0.421         0.394 -0.166  0.157
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
## Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889  1.000
```

As we can see, first 3 components explain nearly 90% of the data scatter, so lets consider the projection of no more than 3 components.

THIRD MODEL

```
c1 <- PC$scores[,1]
c2 <- PC$scores[,2]
c3 <- PC$scores[,3]
Data1 <- data.frame(c1=c1,c2=c2,c3=c3,adi=Data$adi)

model3 <- lm(adi~.-adi, data = Data1[1:(nn-20), ])
summary(model3)
```

```
##
## Call:
## lm(formula = adi ~ . - adi, data = Data1[1:(nn - 20), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1995 -0.7074 -0.1367  0.6052  2.9059
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.41496    0.07181 604.587  < 2e-16 ***
## c1           1.05393    0.03119  33.790  < 2e-16 ***
## c2          -0.25558    0.05360  -4.769 3.68e-06 ***
## c3           1.27968    0.08603  14.874  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9676 on 190 degrees of freedom
## Multiple R-squared:  0.881,  Adjusted R-squared:  0.8791
## F-statistic: 468.9 on 3 and 190 DF,  p-value: < 2.2e-16
```

We got pretty nice result, determination coefficient equals 0.881 and p-value $<$ 2.2e-16 means that there is a dependency between feedback and regressors.

FOURTH MODEL

Consider the model with first three components using 50 sessions before last 20:

```
model4 <- lm(adi~.-adi, data = Data1[(nn-70):(nn-20),])
summary(model4)
```

```
##
## Call:
## lm(formula = adi ~ . - adi, data = Data1[(nn - 70):(nn - 20),
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65791 -0.38810 -0.05871  0.52075  1.18120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.412834   0.687112  63.182  < 2e-16 ***
## c1           1.013889   0.241912   4.191 0.000121 ***
## c2           0.001046   0.212093   0.005 0.996086
## c3           1.438569   0.303490   4.740 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6722 on 47 degrees of freedom
## Multiple R-squared:  0.4597, Adjusted R-squared:  0.4252
## F-statistic: 13.33 on 3 and 47 DF,  p-value: 2.007e-06
```

Result is quite bad, R-squared equals 0.4597, p-value = 2.007e-06, therefore there is a dependency between feedback and regressors.

FIFTH MODEL

Consider model with first component without last 20 sessions:

```r
model5<-lm(adi~.-adi-c2-c3, data = Data1[1:(nn-20), ])
summary(model5)
```

```
##
## Call:
## lm(formula = adi ~ . - adi - c2 - c3, data = Data1[1:(nn - 20),
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5403 -0.8542  0.1561  0.9695  3.4430
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.34010    0.10923  396.79   <2e-16 ***
## c1           1.01134    0.04669   21.66   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.504 on 192 degrees of freedom
## Multiple R-squared:  0.7096, Adjusted R-squared:  0.7081
## F-statistic: 469.3 on 1 and 192 DF,  p-value: < 2.2e-16
```

Determination coefficient (0.7096) is now better than in the last two models, but not really impressive. p-value still shows that there is a dependency between feedback and regressors.

SIXTH MODEL

Lastly, lets create model with first component using 50 sessions before last 20:

```r
model6 <- lm(adi~.-adi-c2-c3, data = Data1[(nn-70):(nn-20), ])
summary(model6)
```

```
##
## Call:
## lm(formula = adi ~ . - adi - c2 - c3, data = Data1[(nn - 70):(nn -
##     20), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.38178 -0.50135  0.03324  0.50646  2.45273
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.0066     0.3849 116.927   <2e-16 ***
## c1            0.2801     0.1642   1.706   0.0944 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8702 on 49 degrees of freedom
## Multiple R-squared:  0.05606,    Adjusted R-squared:  0.03679
## F-statistic:  2.91 on 1 and 49 DF,  p-value: 0.09437
```

Result is pretty damn bad: R-squared is low (0.05606) and p-value shows that there is no dependency.

COMPARISON

Unfortunately, some parts of code just didn't work here (and i have no clue why), so I made another script to compare model6 and model from the previous work.