Week 6 Factor Analysis Assignment

Nitin

October 11, 2020

Step 1 - Data Description

```
library(readxl)
Data <- read_excel("data.xlsx")
```

Data set used for analysis has 25 variables and 361 observation. Among these 25 variables, variables which we are intrested are 13 variables. Variables of our intrest are Month, Coarse wool Price,Copra Price, Cotton Price, Fine wool Price, Hard log Price, Hard sawnwood Price, Hide Price, Plywood Price, Rubber Price, Softlog Price, Soft sawnwood Price, Wood pulp Price .Create separate data frame with 13 variables of our intrest.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## -- Attaching packages ---------------------------------------------------------------------------- t
```

```
## v ggplot2 3.2.0       v purrr   0.2.5
## v tibble  2.1.3       v dplyr   0.8.0.1
## v tidyr   0.8.1       v stringr 1.3.1
## v readr   1.1.1       v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
## -- Conflicts ------------------------------------------------------------------------------- tidyver
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df <- select(Data, -c(3,5,7,9,11,13,15,17,19,21,23,25))
```

```
names(df)<-str_replace_all(names(df), c(" " = "." , "," = "" ))
```

Remove rows will missing data

```
df = na.omit(df)
```

Scale your data

Scale continious verible Month

```
df$Months <- scale(df$Month)
```

```
data_scaled <- as.data.frame(scale(df[,c(2:14)]))
```

```
data_scaled_df <- select(data_scaled, -c(1))
```

Step 2 - Correlation Matrix

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.3
```

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

```
##
## Attaching package: 'car'
```

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

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
model <- lm(Plywood.Price ~., data = data_scaled_df)
vif(model)
```

```
##         Copra.Price        Cotton.Price     Fine.wool.Price
##            5.088118            5.001266            8.103242
##      Hard.log.Price Hard.sawnwood.Price          Hide.Price
##            4.442303            6.360408            1.750596
##        Rubber.Price       Softlog.Price Soft.sawnwood.Price
##            7.346257            2.783009            2.459189
##     Wood.pulp.Price              Months
##            3.160612            4.735338
```

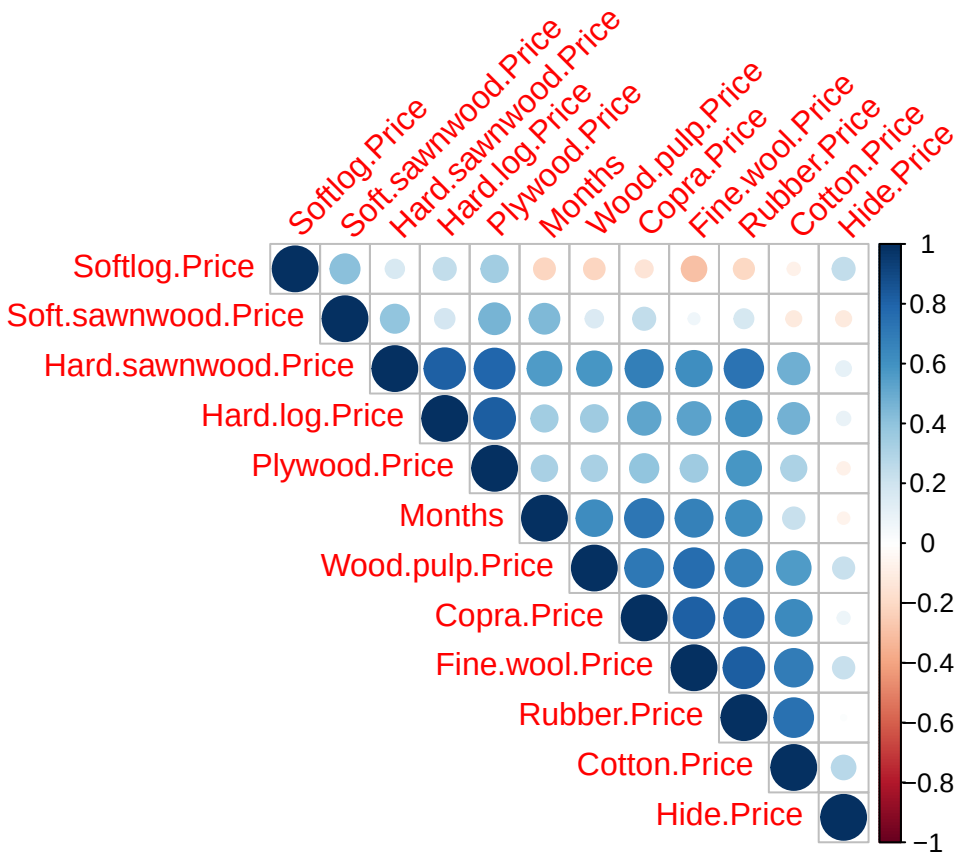
It could be observed that High Variable Inflation Factor (VIF) for variables Copra Price, Cotton Price, Fine wool Price,Hard sawnwood Price, Rubber Price is more then 5.

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.5.3
```

```
## corrplot 0.84 loaded
```

```
cor1 = cor(data_scaled_df)
corrplot(cor1, order = "hclust", type='upper', tl.srt=45)
```



It could be observed that there is strong collinearity between some variables. Plywood Price strongly positively related to Hard sawnwood Price, Hard log Price Rubber Price. On of the variable Hard Sawnwood Price is very strongly positively related to all the variables except Hide Price. Multiple variables strongly corelated to each other can be eliminated and represented just by one of them instead of all the variables while developing model used for multi collinearity analysis.

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.5.3
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
res <- rcorr(as.matrix(cor1))
res$r
```

```
##                     Copra.Price Cotton.Price Fine.wool.Price
```

```
## Copra.Price           1.0000000   0.73007614      0.94601829
## Cotton.Price          0.7300761   1.00000000      0.82949933
## Fine.wool.Price       0.9460183   0.82949933      1.00000000
## Hard.log.Price        0.4294482   0.40711291      0.39156665
## Hard.sawnwood.Price   0.6951658   0.44641773      0.60570748
## Hide.Price           -0.3813097   0.07224627     -0.15842476
## Plywood.Price         0.1846619   0.02776156      0.06668645
## Rubber.Price          0.9282368   0.80580336      0.92666302
## Softlog.Price        -0.8859847  -0.73422799     -0.95338976
## Soft.sawnwood.Price  -0.2379517  -0.72484499     -0.46098473
## Wood.pulp.Price       0.9019613   0.72360859      0.93767356
## Months                0.8545954   0.31906220      0.75937233
##                     Hard.log.Price Hard.sawnwood.Price  Hide.Price
## Copra.Price             0.42944824          0.69516581 -0.38130966
## Cotton.Price            0.40711291          0.44641773  0.07224627
## Fine.wool.Price         0.39156665          0.60570748 -0.15842476
## Hard.log.Price          1.00000000          0.88049904 -0.44732818
## Hard.sawnwood.Price     0.88049904          1.00000000 -0.59480401
## Hide.Price             -0.44732818         -0.59480401  1.00000000
## Plywood.Price           0.86479349          0.79363591 -0.69036213
## Rubber.Price            0.60663627          0.79425848 -0.41740973
## Softlog.Price          -0.14615588         -0.41795906  0.08552014
## Soft.sawnwood.Price    -0.06672517          0.07476851 -0.66251109
## Wood.pulp.Price         0.23534544          0.53570628 -0.15133745
## Months                  0.24069618          0.60319886 -0.56317220
##                     Plywood.Price Rubber.Price Softlog.Price
## Copra.Price            0.18466187    0.9282368   -0.88598466
## Cotton.Price           0.02776156    0.8058034   -0.73422799
## Fine.wool.Price        0.06668645    0.9266630   -0.95338976
## Hard.log.Price         0.86479349    0.6066363   -0.14615588
## Hard.sawnwood.Price    0.79363591    0.7942585   -0.41795906
## Hide.Price            -0.69036213   -0.4174097    0.08552014
## Plywood.Price          1.00000000    0.3793453    0.13693328
## Rubber.Price           0.37934527    1.0000000   -0.83103195
## Softlog.Price          0.13693328   -0.8310319    1.00000000
## Soft.sawnwood.Price    0.38829680   -0.2792230    0.41899369
## Wood.pulp.Price       -0.01229773    0.8422434   -0.93637262
## Months                 0.17710481    0.7443198   -0.78720698
##                     Soft.sawnwood.Price Wood.pulp.Price     Months
## Copra.Price                  -0.23795167      0.90196129  0.8545954
## Cotton.Price                 -0.72484499      0.72360859  0.3190622
## Fine.wool.Price              -0.46098473      0.93767356  0.7593723
## Hard.log.Price               -0.06672517      0.23534544  0.2406962
## Hard.sawnwood.Price           0.07476851      0.53570628  0.6031989
## Hide.Price                   -0.66251109     -0.15133745 -0.5631722
## Plywood.Price                 0.38829680     -0.01229773  0.1771048
## Rubber.Price                 -0.27922296      0.84224340  0.7443198
## Softlog.Price                 0.41899369     -0.93637262 -0.7872070
## Soft.sawnwood.Price           1.00000000     -0.37098676  0.1556571
## Wood.pulp.Price              -0.37098676      1.00000000  0.7799502
## Months                        0.15565710      0.77995017  1.0000000
```

Above is Matrix of Correlation between different variables.

Step 3 - KMO

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:Hmisc':
##
##     describe
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
data_fa <- data_scaled_df[,-7]
matrix_Data <- cor(data_fa)
KMO(r=matrix_Data)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = matrix_Data)
## Overall MSA =  0.76
## MSA for each item =
##         Copra.Price        Cotton.Price     Fine.wool.Price
##                0.87                0.71                0.82
##      Hard.log.Price Hard.sawnwood.Price          Hide.Price
##                0.70                0.78                0.32
##        Rubber.Price       Softlog.Price Soft.sawnwood.Price
##                0.82                0.42                0.54
##     Wood.pulp.Price              Months
##                0.90                0.78
```

It can be observed that MSA is 0.76 which is greater than 0.5 therefore factor analysis is appropriate on this data.

Step 4 - Number of Factors

```
ev <- eigen(cor(data_fa))
ev$values
```

```
##  [1] 5.49562152 1.72399106 1.40725501 0.88944392 0.46675754 0.29286481
##  [7] 0.26096254 0.16672309 0.14096498 0.08749761 0.06791791
```

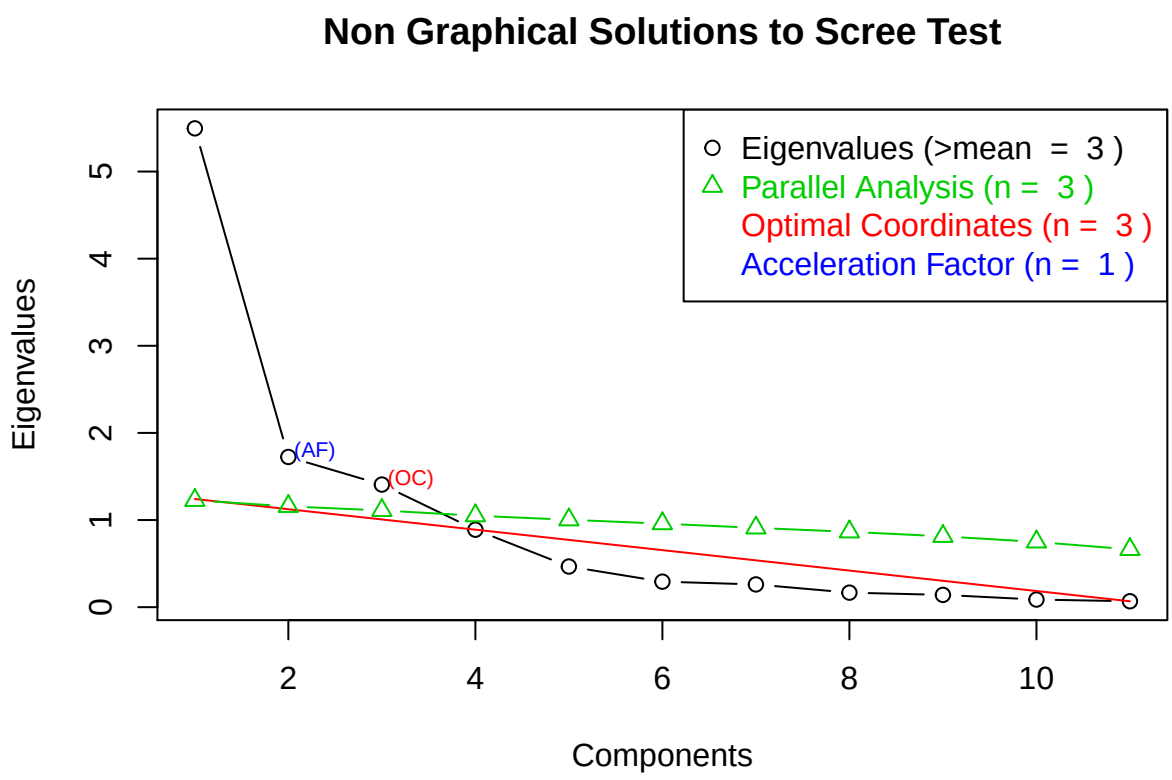
```
library(nFactors)
```

```
## Warning: package 'nFactors' was built under R version 3.5.3
```

```
##
## Attaching package: 'nFactors'
```

```
## The following object is masked from 'package:lattice':
##
##     parallel
```

```
app <- parallel(subject=nrow(data_scaled_df),var=ncol(data_fa),
  rep=100,cent=.05)
nS <- nScree(x=ev$values, aparallel=app$eigen$qevpea)
plotnScree(nS)
```



Above plot display Eigenvalues for Components, which are helpful in determining number of factors.

Step 5 - Run Analysis

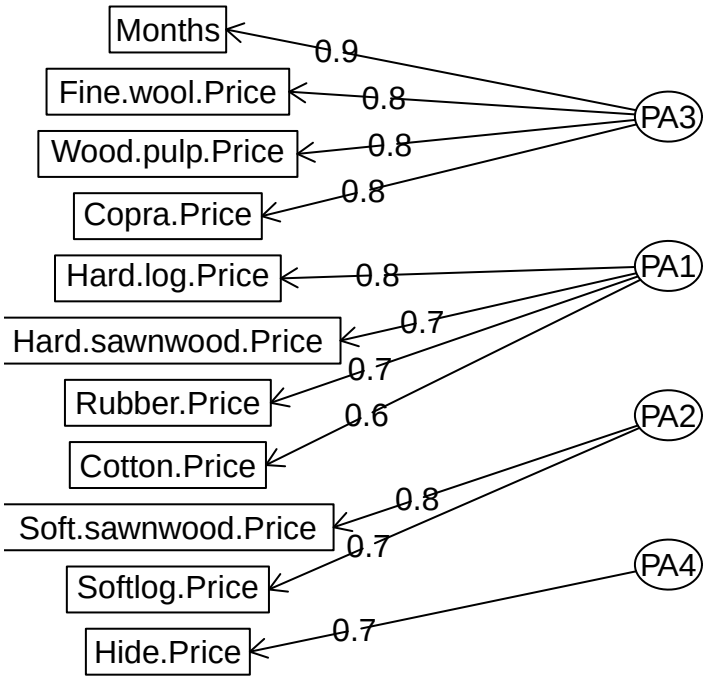
```
nfactors <- 4
Anyl <-factanal(data_fa,nfactors,scores = c("regression"),rotation = "varimax")
print(Anyl)
```

```
##
## Call:
## factanal(x = data_fa, factors = nfactors, scores = c("regression"),     rotation = "varimax")
##
## Uniquenesses:
##         Copra.Price        Cotton.Price     Fine.wool.Price
##               0.191               0.056               0.149
##      Hard.log.Price Hard.sawnwood.Price          Hide.Price
##               0.005               0.190               0.810
##        Rubber.Price       Softlog.Price Soft.sawnwood.Price
##               0.167               0.005               0.427
##     Wood.pulp.Price              Months
##               0.345               0.109
```

```
##
## Loadings:
##                     Factor1 Factor2 Factor3 Factor4
## Copra.Price          0.843   0.262   0.123   0.120
## Cotton.Price         0.627   0.263  -0.323   0.614
## Fine.wool.Price      0.869   0.284
## Hard.log.Price       0.300   0.940           0.109
## Hard.sawnwood.Price  0.551   0.652   0.268
## Hide.Price                                   0.430
## Rubber.Price         0.808   0.387           0.159
## Softlog.Price       -0.429   0.266   0.663   0.549
## Soft.sawnwood.Price  0.176           0.732
## Wood.pulp.Price      0.793   0.105           0.109
## Months               0.805   0.101   0.389  -0.286
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings        4.4   1.776   1.343   1.027
## Proportion Var     0.4   0.161   0.122   0.093
## Cumulative Var     0.4   0.561   0.684   0.777
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 254.24 on 17 degrees of freedom.
## The p-value is 2.83e-44
```

In can be observeds that we have 777 % data variance explained with 4 factors.

```
favar <-  fa(r=data_fa, nfactors = 4, rotate="varimax",fm="pa")
fa.diagram(favar)
```



Highly corelated variables are grouped together using factor analysis.

```
head(favar$scores)
```

```
##              PA3        PA1       PA2      PA4
## [1,]  0.06516367 -0.9090737 -2.558568 1.821247
## [2,] -0.01037673 -0.9621974 -2.633884 1.858806
## [3,] -0.32401226 -0.8154397 -2.533951 1.598135
## [4,] -0.33090509 -0.8179214 -2.602968 1.335279
## [5,] -0.28825290 -0.8043492 -2.451381 1.450050
## [6,] -0.33430499 -0.7480897 -2.490981 1.360564
```

```
regdata <- cbind(data_scaled_df[7], favar$scores)
```

Labeling the data

lable factors aptoprestly as Ingredient, Time, Softlog, Hide.

```
names(regdata) <- c("Plywood_Price", "Ingredient", "Time", "Softlog", "Hide"  )
head(regdata)
```

```
##   Plywood_Price  Ingredient       Time   Softlog     Hide
## 1     -2.103453  0.06516367 -0.9090737 -2.558568 1.821247
## 2     -1.700402 -0.01037673 -0.9621974 -2.633884 1.858806
## 3     -1.446147 -0.32401226 -0.8154397 -2.533951 1.598135
## 4     -1.397686 -0.33090509 -0.8179214 -2.602968 1.335279
## 5     -1.545842 -0.28825290 -0.8043492 -2.451381 1.450050
## 6     -1.328946 -0.33430499 -0.7480897 -2.490981 1.360564
```

Step 6 - Run Regression

```
set.seed(100)
model1 = lm(Plywood_Price~.,regdata)
summary(model1)
```

```
##
## Call:
## lm(formula = Plywood_Price ~ ., data = regdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30204 -0.29526 -0.04154  0.26359  1.43279
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.820e-16  2.512e-02   0.000  1.00000
## Ingredient   7.531e-02  2.637e-02   2.856  0.00457 **
## Time         8.428e-01  2.743e-02  30.726  < 2e-16 ***
## Softlog      4.100e-01  2.688e-02  15.251  < 2e-16 ***
## Hide        -1.635e-01  2.891e-02  -5.654 3.45e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4543 on 322 degrees of freedom
## Multiple R-squared:  0.7962, Adjusted R-squared:  0.7936
## F-statistic: 314.4 on 4 and 322 DF,  p-value: < 2.2e-16
```

It could be observed that Pvalues for all variables are significant. R-squares is 0.7962 and adjusted R-squared is 0.7936.