BUDA535Module2

## Problem 1 :

The SMarket data in the ISLR package describes 1250 days of daily percentage returns for the S&P between 2001 and 2005. Use factor analysis and principal components analysis to try and find relevant summary statistics that could provide insights into the data. Report your findings including insights into how you selected your number of statistics.

library(ISLR)  
data("Smarket")  
#head(Smarket)  
  
Smarket$Direction <- ifelse(Smarket$Direction == "Up", 1, 0)  
Smarket\_newdata <- Smarket[]  
#head(Smarket\_newdata)  
summary(Smarket\_newdata)

## Year Lag1 Lag2   
## Min. :2001 Min. :-4.922000 Min. :-4.922000   
## 1st Qu.:2002 1st Qu.:-0.639500 1st Qu.:-0.639500   
## Median :2003 Median : 0.039000 Median : 0.039000   
## Mean :2003 Mean : 0.003834 Mean : 0.003919   
## 3rd Qu.:2004 3rd Qu.: 0.596750 3rd Qu.: 0.596750   
## Max. :2005 Max. : 5.733000 Max. : 5.733000   
## Lag3 Lag4 Lag5   
## Min. :-4.922000 Min. :-4.922000 Min. :-4.92200   
## 1st Qu.:-0.640000 1st Qu.:-0.640000 1st Qu.:-0.64000   
## Median : 0.038500 Median : 0.038500 Median : 0.03850   
## Mean : 0.001716 Mean : 0.001636 Mean : 0.00561   
## 3rd Qu.: 0.596750 3rd Qu.: 0.596750 3rd Qu.: 0.59700   
## Max. : 5.733000 Max. : 5.733000 Max. : 5.73300   
## Volume Today Direction   
## Min. :0.3561 Min. :-4.922000 Min. :0.0000   
## 1st Qu.:1.2574 1st Qu.:-0.639500 1st Qu.:0.0000   
## Median :1.4229 Median : 0.038500 Median :1.0000   
## Mean :1.4783 Mean : 0.003138 Mean :0.5184   
## 3rd Qu.:1.6417 3rd Qu.: 0.596750 3rd Qu.:1.0000   
## Max. :3.1525 Max. : 5.733000 Max. :1.0000

cor(Smarket\_newdata[,])

## Year Lag1 Lag2 Lag3 Lag4  
## Year 1.00000000 0.029699649 0.030596422 0.033194581 0.035688718  
## Lag1 0.02969965 1.000000000 -0.026294328 -0.010803402 -0.002985911  
## Lag2 0.03059642 -0.026294328 1.000000000 -0.025896670 -0.010853533  
## Lag3 0.03319458 -0.010803402 -0.025896670 1.000000000 -0.024051036  
## Lag4 0.03568872 -0.002985911 -0.010853533 -0.024051036 1.000000000  
## Lag5 0.02978799 -0.005674606 -0.003557949 -0.018808338 -0.027083641  
## Volume 0.53900647 0.040909908 -0.043383215 -0.041823686 -0.048414246  
## Today 0.03009523 -0.026155045 -0.010250033 -0.002447647 -0.006899527  
## Direction 0.07460829 -0.039757295 -0.024081472 0.006131513 0.004215206  
## Lag5 Volume Today Direction  
## Year 0.029787995 0.53900647 0.030095229 0.074608288  
## Lag1 -0.005674606 0.04090991 -0.026155045 -0.039757295  
## Lag2 -0.003557949 -0.04338321 -0.010250033 -0.024081472  
## Lag3 -0.018808338 -0.04182369 -0.002447647 0.006131513  
## Lag4 -0.027083641 -0.04841425 -0.006899527 0.004215206  
## Lag5 1.000000000 -0.02200231 -0.034860083 0.005422568  
## Volume -0.022002315 1.00000000 0.014591823 0.022950959  
## Today -0.034860083 0.01459182 1.000000000 0.730562902  
## Direction 0.005422568 0.02295096 0.730562902 1.000000000

cor(Smarket)

## Year Lag1 Lag2 Lag3 Lag4  
## Year 1.00000000 0.029699649 0.030596422 0.033194581 0.035688718  
## Lag1 0.02969965 1.000000000 -0.026294328 -0.010803402 -0.002985911  
## Lag2 0.03059642 -0.026294328 1.000000000 -0.025896670 -0.010853533  
## Lag3 0.03319458 -0.010803402 -0.025896670 1.000000000 -0.024051036  
## Lag4 0.03568872 -0.002985911 -0.010853533 -0.024051036 1.000000000  
## Lag5 0.02978799 -0.005674606 -0.003557949 -0.018808338 -0.027083641  
## Volume 0.53900647 0.040909908 -0.043383215 -0.041823686 -0.048414246  
## Today 0.03009523 -0.026155045 -0.010250033 -0.002447647 -0.006899527  
## Direction 0.07460829 -0.039757295 -0.024081472 0.006131513 0.004215206  
## Lag5 Volume Today Direction  
## Year 0.029787995 0.53900647 0.030095229 0.074608288  
## Lag1 -0.005674606 0.04090991 -0.026155045 -0.039757295  
## Lag2 -0.003557949 -0.04338321 -0.010250033 -0.024081472  
## Lag3 -0.018808338 -0.04182369 -0.002447647 0.006131513  
## Lag4 -0.027083641 -0.04841425 -0.006899527 0.004215206  
## Lag5 1.000000000 -0.02200231 -0.034860083 0.005422568  
## Volume -0.022002315 1.00000000 0.014591823 0.022950959  
## Today -0.034860083 0.01459182 1.000000000 0.730562902  
## Direction 0.005422568 0.02295096 0.730562902 1.000000000

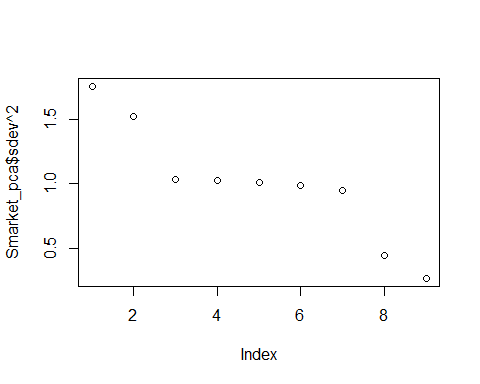
Smarket\_pca <- prcomp(Smarket\_newdata,scale = TRUE)  
Smarket\_pca

## Standard deviations (1, .., p=9):  
## [1] 1.3255852 1.2344741 1.0171453 1.0144396 1.0066004 0.9927133 0.9746893  
## [8] 0.6637594 0.5156601  
##   
## Rotation (n x k) = (9 x 9):  
## PC1 PC2 PC3 PC4  
## Year -0.230516613 0.661620498 -0.0551187254 -0.015459168  
## Lag1 0.037695706 0.117063240 0.3855593444 -0.108092110  
## Lag2 0.031297066 -0.009004774 -0.6879026427 -0.079213831  
## Lag3 -0.004748154 -0.015406193 0.4903224940 0.417813790  
## Lag4 0.002867951 -0.017107141 0.0650032984 -0.767531609  
## Lag5 0.022200343 0.022218525 -0.3610799861 0.466742879  
## Volume -0.197812789 0.677622766 0.0008015333 0.007829558  
## Today -0.668573360 -0.223287665 0.0016988059 -0.009243211  
## Direction -0.676622314 -0.196009609 -0.0084859022 0.011553815  
## PC5 PC6 PC7 PC8 PC9  
## Year -0.100780837 -0.088545450 0.09147339 0.688040518 0.077968834  
## Lag1 0.540407382 0.508231724 0.52342186 0.005093384 -0.014117649  
## Lag2 -0.357376535 0.402594050 0.46860976 -0.098352367 -0.019654045  
## Lag3 -0.568409909 -0.161608946 0.47374871 -0.105385480 0.002091189  
## Lag4 -0.047210790 -0.517926187 0.35011948 -0.115270133 0.004655732  
## Lag5 0.492089996 -0.523056716 0.35729720 -0.078947224 0.033925836  
## Volume -0.008305265 0.022085663 -0.14371507 -0.691459459 -0.048047734  
## Today 0.030303564 0.057593714 0.02741306 -0.080598273 0.701128418  
## Direction 0.040330708 -0.004485763 0.04448358 0.040778454 -0.705881447

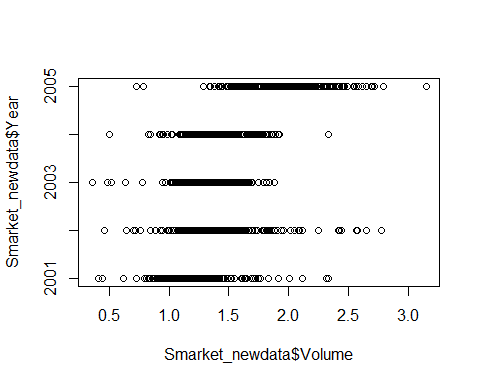
summary(Smarket\_pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.3256 1.2345 1.0171 1.0144 1.0066 0.9927 0.9747  
## Proportion of Variance 0.1952 0.1693 0.1149 0.1143 0.1126 0.1095 0.1056  
## Cumulative Proportion 0.1952 0.3646 0.4795 0.5939 0.7065 0.8159 0.9215  
## PC8 PC9  
## Standard deviation 0.66376 0.51566  
## Proportion of Variance 0.04895 0.02955  
## Cumulative Proportion 0.97045 1.00000

plot(Smarket\_pca$sdev^2)



plot(Smarket\_newdata$Year ~ Smarket\_newdata$Volume)



g1<- scale(Smarket\_newdata)%\*%Smarket\_pca$rotation[,]  
#plot(data.frame(g1))  
  
fs1<-factanal(scale(Smarket\_newdata),factors = 3)  
fs1

##   
## Call:  
## factanal(x = scale(Smarket\_newdata), factors = 3)  
##   
## Uniquenesses:  
## Year Lag1 Lag2 Lag3 Lag4 Lag5 Volume   
## 0.005 0.997 0.993 0.994 0.992 0.997 0.005   
## Today Direction   
## 0.463 0.005   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## Year 0.877 0.470   
## Lag1   
## Lag2   
## Lag3   
## Lag4   
## Lag5   
## Volume 0.873 -0.483   
## Today 0.732   
## Direction 0.997   
##   
## Factor1 Factor2 Factor3  
## SS loadings 1.538 1.533 0.478  
## Proportion Var 0.171 0.170 0.053  
## Cumulative Var 0.171 0.341 0.394  
##   
## Test of the hypothesis that 3 factors are sufficient.  
## The chi square statistic is 10.44 on 12 degrees of freedom.  
## The p-value is 0.578

fs2<-factanal(scale(Smarket\_newdata),factors = 4)  
fs2

##   
## Call:  
## factanal(x = scale(Smarket\_newdata), factors = 4)  
##   
## Uniquenesses:  
## Year Lag1 Lag2 Lag3 Lag4 Lag5 Volume   
## 0.005 0.996 0.993 0.994 0.992 0.994 0.005   
## Today Direction   
## 0.005 0.005   
##   
## Loadings:  
## Factor1 Factor2 Factor3 Factor4  
## Year 0.898 0.406 -0.130   
## Lag1   
## Lag2   
## Lag3   
## Lag4   
## Lag5   
## Volume 0.846 -0.527   
## Today 0.801 0.593   
## Direction 0.990 -0.101   
##   
## Factor1 Factor2 Factor3 Factor4  
## SS loadings 1.631 1.525 0.470 0.385  
## Proportion Var 0.181 0.169 0.052 0.043  
## Cumulative Var 0.181 0.351 0.403 0.446  
##   
## Test of the hypothesis that 4 factors are sufficient.  
## The chi square statistic is 6.15 on 6 degrees of freedom.  
## The p-value is 0.406

### Interpretation :

**Data :** Used all the variables for PCA and factor analysis. Converted factor variable into binary 0 and 1. As per the summary of data, variables lag1, lag2, lag3, lag4, lag5 and today have similar Min, max, median and Mean.

**Principal Component Analysis :** As per the scree plot, the elbow happens close to 3 and from 3 to 7, variance is almost similar. Deciding the number of components that takes around 80% of threshold and looking at standard deviation of PCs that start going below 1, PC1 to PC5 can be considered.

PC1 is dominated by variables today and direction. Percentage return today is indicated by direction variable if its positive or negative return.

PC2 is dominated by variables year and volume. Volume of shares traded increases with the year.

PC3, PC4 and PC5 are inter related with Lag1 to lag5.

Overall, PC1 and PC2 gives more meaningful insights. Plot of Volume~Year shows the similar insight. PC3 to PC5 is more on percentage return on previous days 1 to 5. Percentage return insights can be more discussed with expertise and can futher be drilled down and can see if PCA is good for this dataset or not.

**Factor Analysis :** 3 and 4 factors model created. Choosing 3 factors analysis as p-value is above .05 which says we fail to reject the null hypothesis. Hence it says that we are using a good approximation.

Factor 1 shows average of today and direction. Factor 2 shows overall average of Volume and year. Factor 3 is the differences of volume and year. A high uniqueness for a variable indicates that the factors do not account well for its variance. Lag1 thru lag4 variables have high uniqueness. Factor analysis shows few similar insights like PCA.

## Problem 2 :

The Anscombe data in the car package describes US Public School Expenditures in the 1970’s for the 50 states plus Washington DC. Use PCA and Factor analysis to see if there are possible ways to gain insights on the relationships between the 4 variables, or produce summary statistics that could provide intuition about spending. Report your findings and comment on how you came up with decisions.

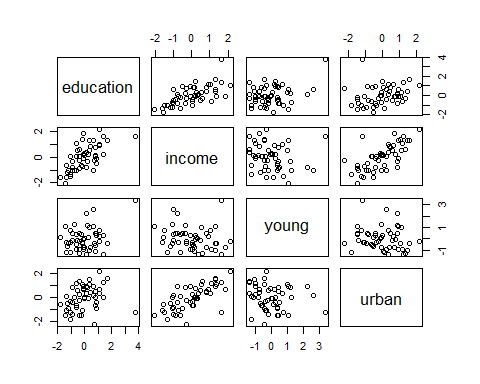
library(car)

## Loading required package: carData

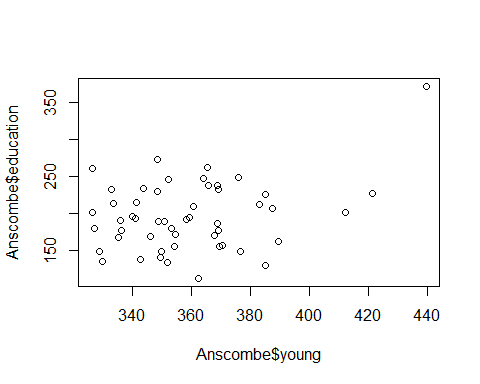
data(Anscombe)  
head(Anscombe)

## education income young urban  
## ME 189 2824 350.7 508  
## NH 169 3259 345.9 564  
## VT 230 3072 348.5 322  
## MA 168 3835 335.3 846  
## RI 180 3549 327.1 871  
## CT 193 4256 341.0 774

plot(as.data.frame(scale(Anscombe)))



plot(Anscombe$education~Anscombe$young)



cor(Anscombe)

## education income young urban  
## education 1.0000000 0.6675773 0.3114855 0.2633238  
## income 0.6675773 1.0000000 -0.1623600 0.6854580  
## young 0.3114855 -0.1623600 1.0000000 -0.1386334  
## urban 0.2633238 0.6854580 -0.1386334 1.0000000

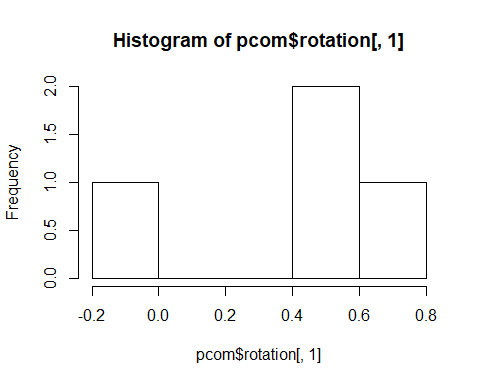
pcom <- prcomp(Anscombe,scale = TRUE)  
pcom

## Standard deviations (1, .., p=4):  
## [1] 1.4483585 1.1139820 0.7412562 0.3344265  
##   
## Rotation (n x k) = (4 x 4):  
## PC1 PC2 PC3 PC4  
## education 0.52501731 -0.4592067 0.4843456 -0.5281055  
## income 0.65835928 0.1092565 0.1783350 0.7230648  
## young -0.01648445 -0.8331116 -0.4873262 0.2610870  
## urban 0.53911793 0.2883006 -0.7043572 -0.3607152

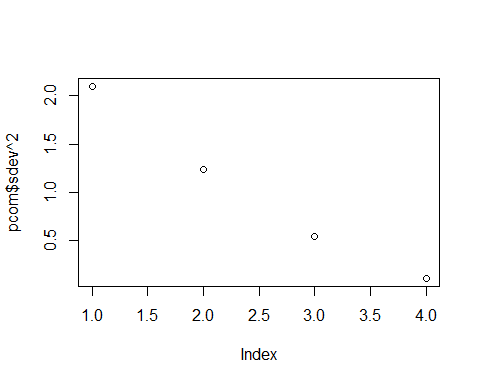
summary(pcom)

## Importance of components:  
## PC1 PC2 PC3 PC4  
## Standard deviation 1.4484 1.1140 0.7413 0.33443  
## Proportion of Variance 0.5244 0.3102 0.1374 0.02796  
## Cumulative Proportion 0.5244 0.8347 0.9720 1.00000

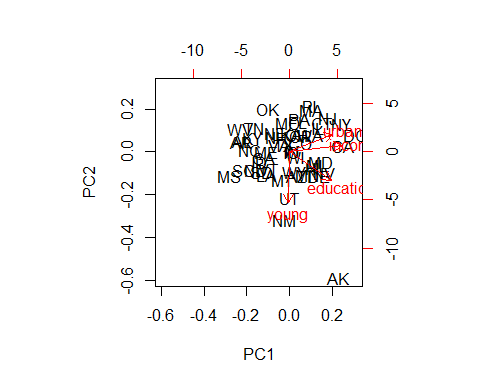
hist(pcom$rotation[,1])



plot(pcom$sdev^2)



biplot(pcom)



fs<-factanal(scale(Anscombe),factors = 1)  
fs

##   
## Call:  
## factanal(x = scale(Anscombe), factors = 1)  
##   
## Uniquenesses:  
## education income young urban   
## 0.554 0.005 0.974 0.529   
##   
## Loadings:  
## Factor1  
## education 0.668   
## income 0.998   
## young -0.160   
## urban 0.686   
##   
## Factor1  
## SS loadings 1.937  
## Proportion Var 0.484  
##   
## Test of the hypothesis that 1 factor is sufficient.  
## The chi square statistic is 27.61 on 2 degrees of freedom.  
## The p-value is 1.01e-06

### Interpretation :

**Principal Component Analysis :** Based on the summary and scree plot, PC1 and PC2 gives around 85% of variabilty. PC1 is average of all three variables : income, education and urban versus young. Meaning, in urban area, education cost and income are higher and all these 3 variables are related with each other and young population under 18 is less hence school education cost will be less.

PC2 is dominated by variables young and education cost with same magnitude (minus) hence it will be sum with negative sign. young and education versus income and urban. Meaning, school education cost will be high with young population under 18 years in the area. This will be sum(income,urban) - sum(young,education). young people under 18 are not likely to be located in downtowns/core urban areas.

biplot shows the similar insights for PC1 and PC2.

**Factor Analysis :** Allowed to create 1 factor model as variables are too less. p-value is < 0.05 which says we reject the null hypothesis. Which says we might not be using good approximation?

Factor 1 shows overall sum of education, income and urban variables minus young. It gives the same insight as PC1 of PCA. A high uniqueness for a variable (young) indicates that the factors do not account well for its variance.