NAME: Omotayo Abdul-Hakeem

CLASS: STA 135

INSTRUCTOR'S NAME: Xiaodong Li

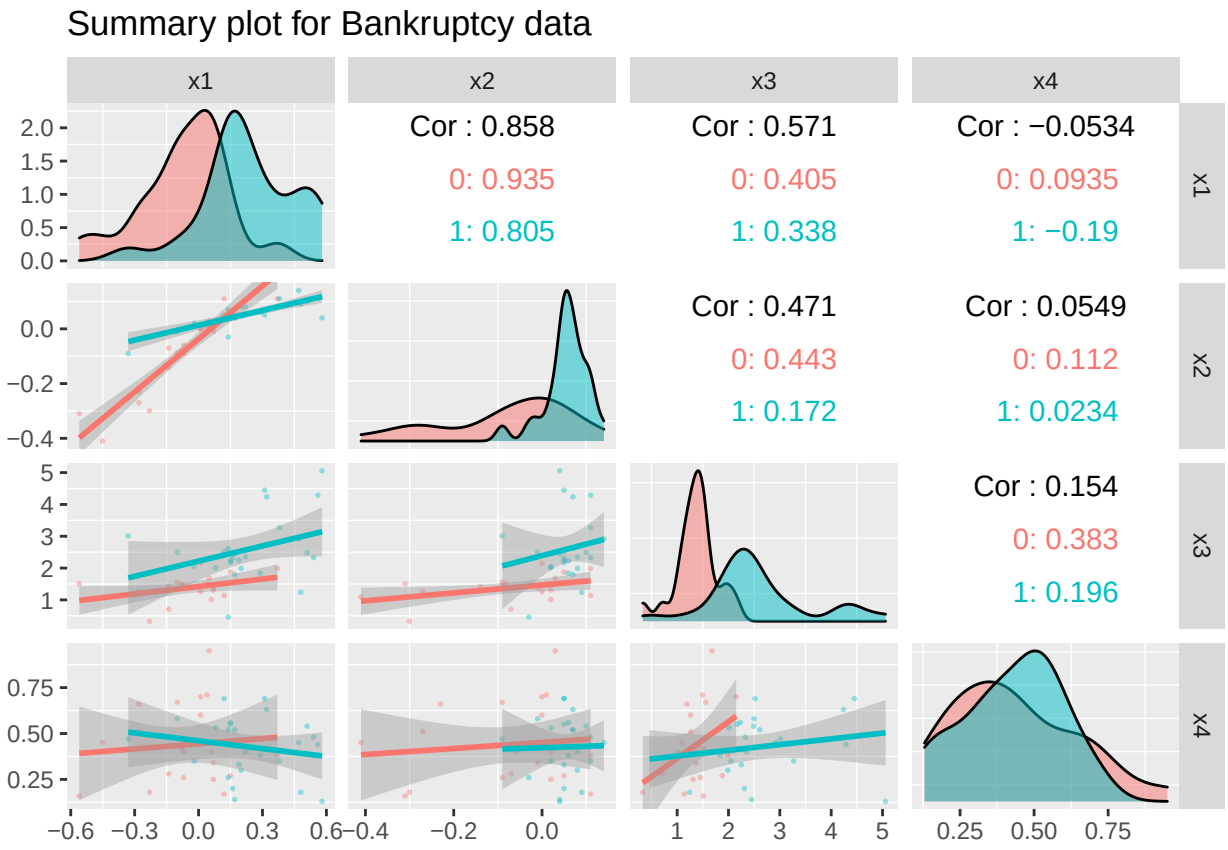
Dataset 1 - Two Sample T-Test and Linear Discrimant Matrix

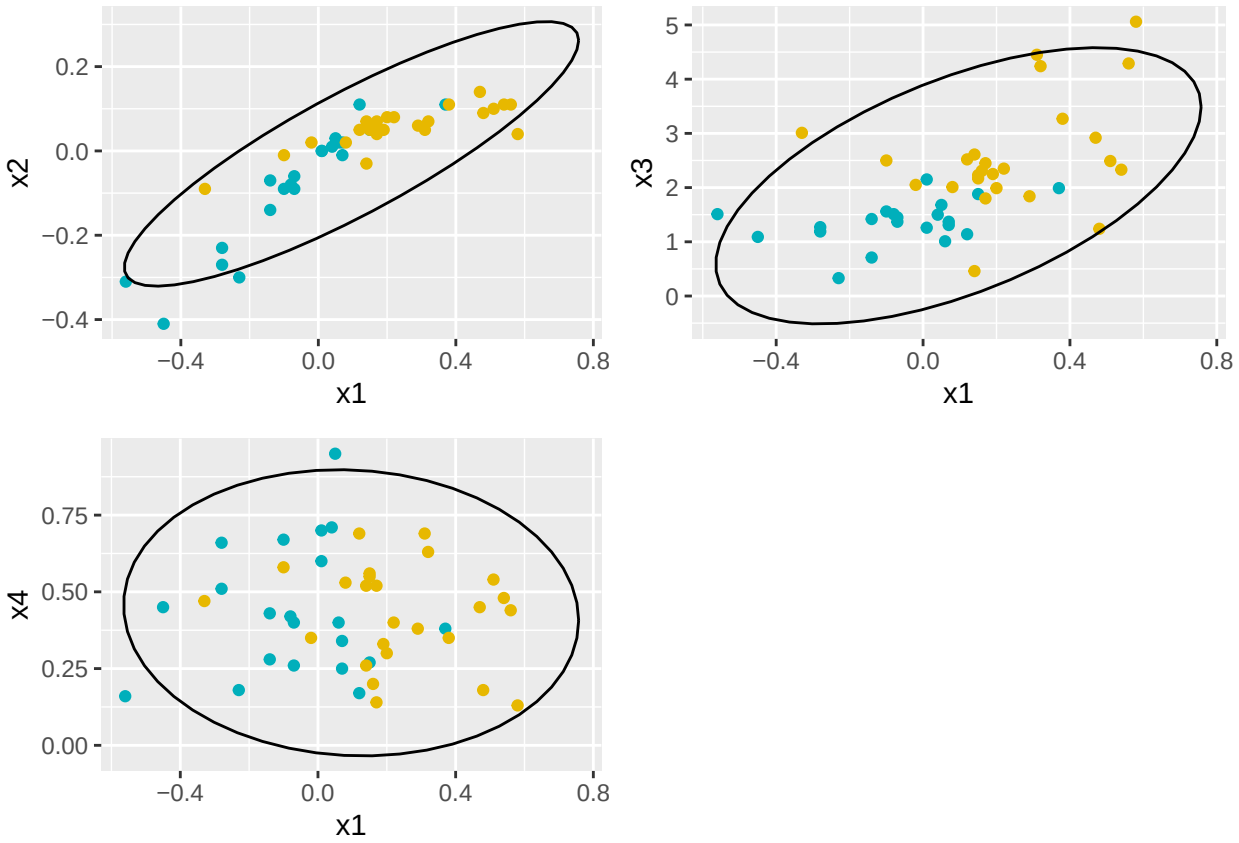
Introduction

In this report, we analyse the annual financial data collected for bankrupt firms approximately 2 years prior to their bankruptcy and for financially sound firms at about the same time. We aim to perfor Linear discriminat analysis for classifying firms into bankruptcy or nonbankrupcy. Also we would perform a two-sample t-test to see if there is a significant difference in the mean of bankrupt and non-bankrupt firms. The source of the dataset is from Applied Multivariate Statistical Analysis, 5th Edition as T11_04.DAT

Summary

In our dataset we have 46 firms with 21 being bakrupt and 25 being no bankrupt. The dataset has four variables, X_1 = (cash flow)/ (total debt),X_2 = NI/TA =(net income)j(total assets), X_3 = CA/CL =(current assets)/(current liabilities), and X_4 = CA/NS =(current assets)/(net sales). From the summary plot, we see that x2 and x1 are highly correlated and likewise x2 and x3. The distributions of the variables are largely normal. For x1 and x4, both population have similar distibution while for x2 and x3, the distributions are unequal.





Analysis

First, we plot the data for pairs of observations (x1, x2), (x1,x3), (x1,x4). From the plot, we see that the shaps in each plots are fairly elliptical. Thus, we can assume the variables follow bivariate normal distibution. For this analysi, we performed LDA of two sets of variable at a time with equal prior. We label the bankrupt population as Group 1 and the nonbankrupt population as Group 2

LDA Analysis

1) x1 and x2: We presnt the result below. The model has 0.239 as the apparent error rate (APER)

Table 1: Table 1: LDA for x1 and x2

	x1	x2
Group 1 mean	-0.069	-0.081
Group 2 mean	0.235	0.055
LDA coefficients	3.271	3.367

2) x1 and x3: We presnt the result below. The model has 0.13 as the apparent error rate (APER)

Table 2: Table 2: LDA for x1 and x3

	x1	x3
Group 1 mean	-0.069	-0.081
Group 2 mean	0.235	0.055

	x1	x3
LDA coefficients	2.664	3.367

3) x1 and x4: We presnt the result below. The model has 0.196 as the apparent error rate (APER)

Table 3: Table 3: LDA for x1 and x4

	x1	x4
Group 1 mean	-0.0690	-0.081
Group 2 mean	0.2350	0.055
LDA coefficients	4.6773	3.367

4) x2 and x3: We presnt the result below. The model has 0.13 as the apparent error rate (APER)

Table 4: Table 4: LDA for x2 and x3

	x1	x4
Group 1 mean	-0.0690	-0.081
Group 2 mean	0.2350	0.055
LDA coefficients	4.6773	5.496

5) x2 and x4: We presnt the result below. The model has 0.239 as the apparent error rate (APER)

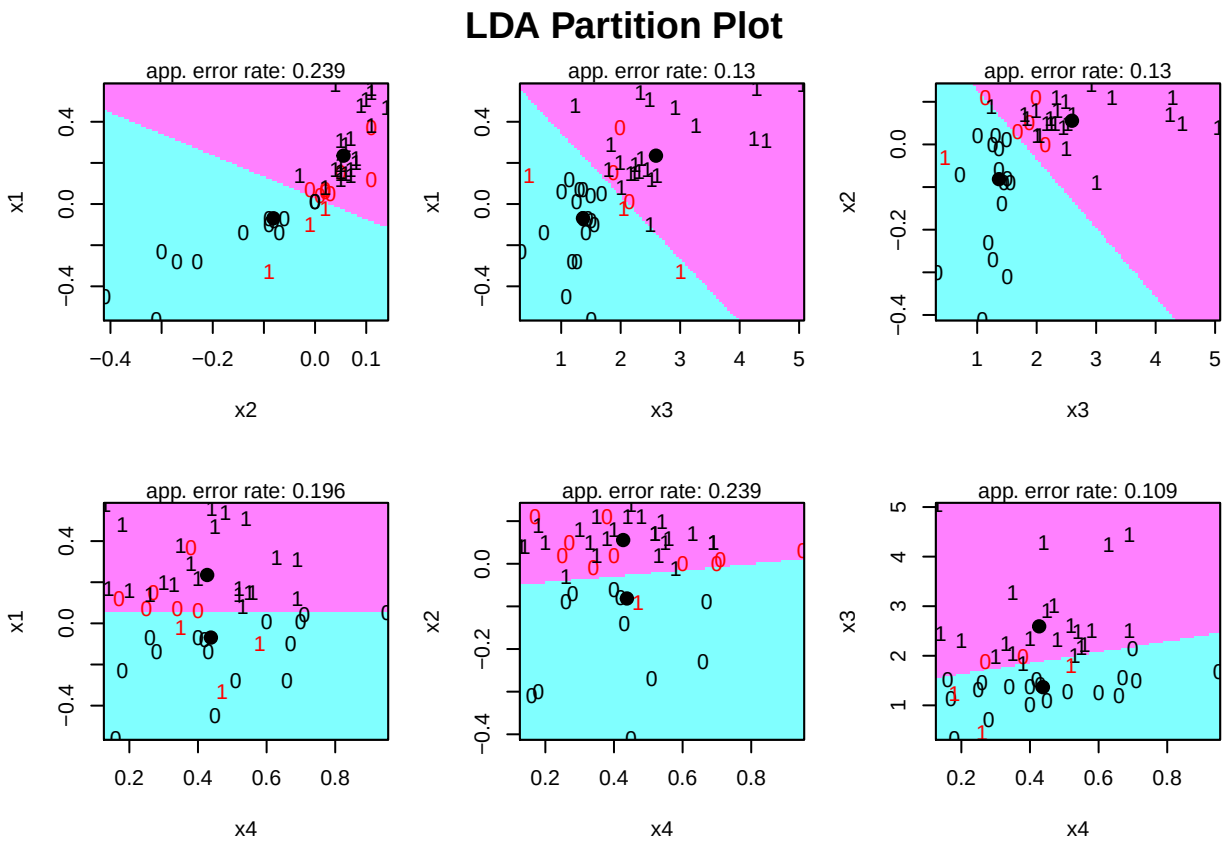
Table 5: Table 5: LDA for x2 and x4

	x2	x4
Group 1 mean	-0.0690	-0.081
Group 2 mean	0.2350	0.055
LDA coefficients	4.6773	9.630

6) x3 and x4: We presnt the result below. The model has 0.109 as the apparent error rate (APER)

Table 6: Table 6: LDA for x3 and x4

	x3	x4
Group 1 mean	1.3661	0.4376
Group 2 mean	2.5936	0.4268
LDA coefficients	1.2745	-1.4000



7) We now consider the model with variable x1, x2, x3, and x4.

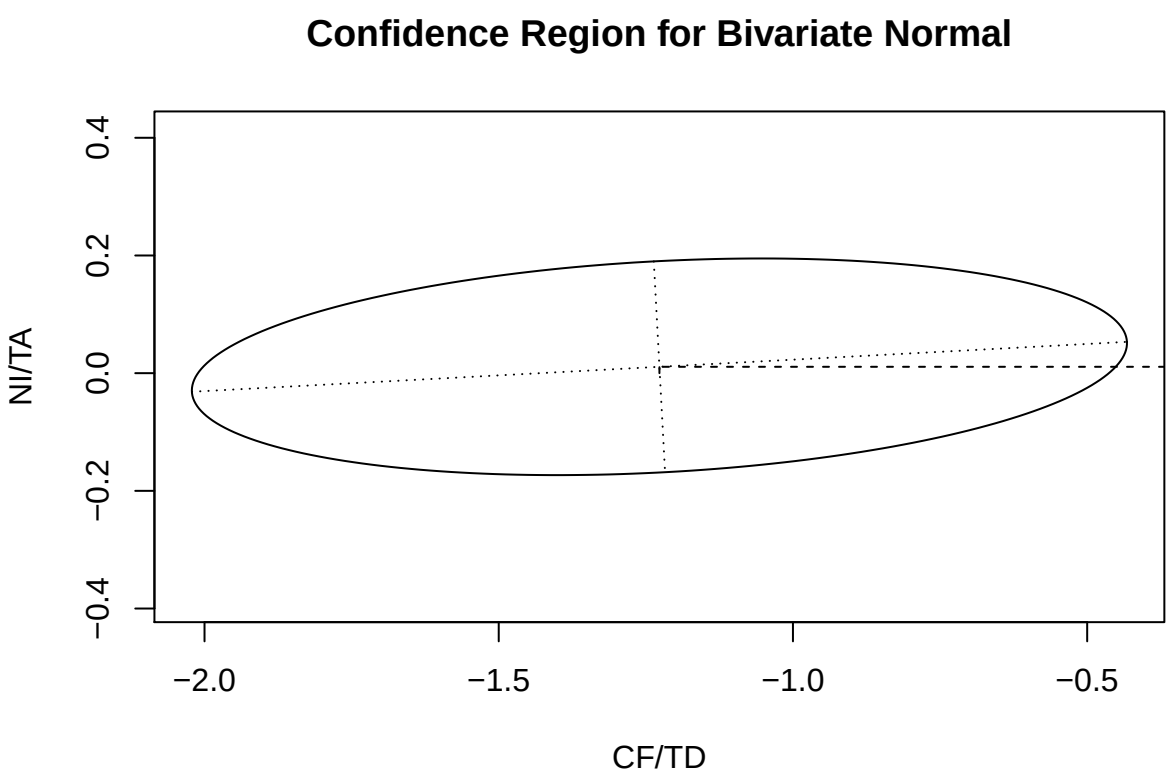
Table 7: Table 6: LDA for x1, x2, x3 and x4

	x1	x2	x3	x4
Group 1 mean	-0.0690	-0.0810	1.3661	0.4376
Group 2 mean	0.2350	0.0550	2.5936	0.4268
LDA coefficients	0.6612	4.3935	0.8872	-1.1785
### T-test analysis				

For the two sample t-test analysis, we consider just only two variables. x3 and x4 were the two chosen variables due to correlation found the summary section.

We test for the hypothesis $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$

We the reject the null hypothesis since the test statistic is greater than the critical value.(ie $28.45 > 11.17$)



Since we reject the null hypothesis, we use simultaneous confidence intervals to check the significant components.

For x3: $[-2.021, -0.4323]$ For x4: $[-0.173, 0.1949]$

Then we compute the Bonferroni simultaneous confidence intervals:

For x3: $[-1.8464, -0.6074]$ For x4: $[-0.1326, 0.154]$

Conclusion

In our LDA, we computed discriminant function for all possible combinations of two variables. We notice that the LDA model with x3 and x4 has the best apparent error rate of all with 0.109. In the two sample T-test analysis, we used reduce our dataset to just two variables x3 and x4. We tested for equality of means and reject the hypothesis that the means for the bankrupt and nonbankrupt group have the same mean. We also compute the simultaneous confidence intervals to investigate where the difference lies.

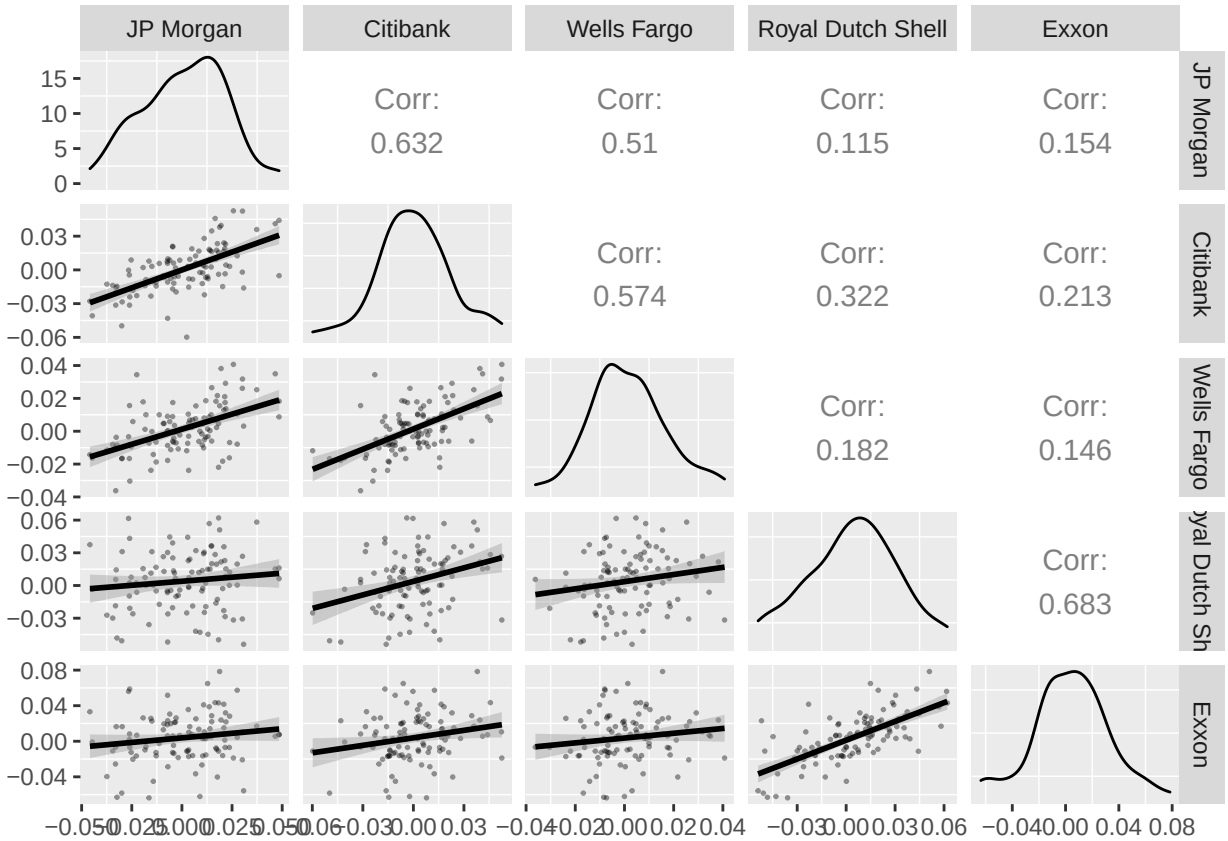
DATASET 2 - Principal Component Analysis

Introduction

The weekly rates of return for five stocks listed on the New York Stock Exchange are analyzed in this report. The Stock-price data consists of 103 weekly rates of return on 5 stocks. The weekly rates of return for five stocks (JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell, and ExxonMobil) listed on the New York Stock Exchange were determined for the period January 2004 through December 2005. The weekly rates of return are defined as (current week closing price-previous week closing price)/(previous week closing price), adjusted for stock splits and dividends. The source of the dataset is from Applied Multivariate Statistical Analysis, 5th Edition as T08_04.DAT

Summary

We notice ther is linear correlation anonges all stocks. Citi bank, JP Morgan and Wells Fargo stocks are highly correlated.



Analysis

For the principal component analysis, we start with computing the covariance matrix.

Table 1: Covariance matrix Forbes Dataset

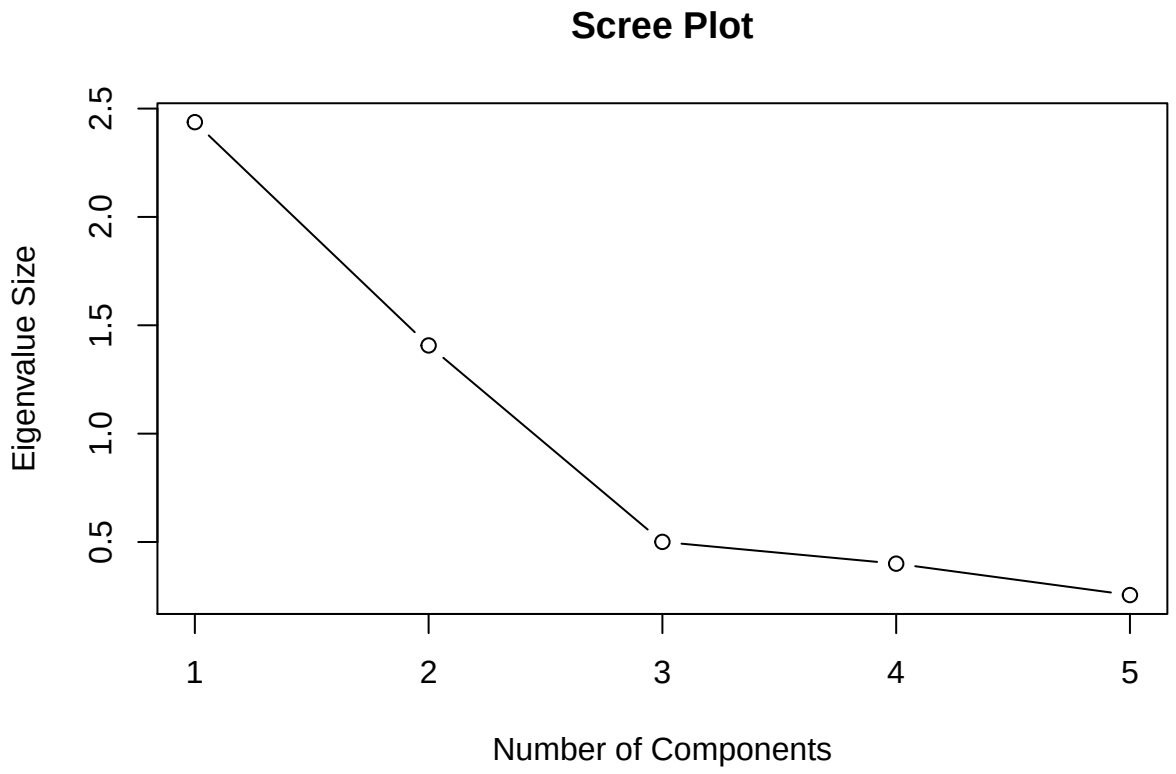
	JP.Morgan	Citibank	Wells.Fargo	Royal.Dutch.Shell	Exxon
JP Morgan	0.0004333	0.0002757	1.59e-04	0.0000641	0.0000890
Citibank	0.0002757	0.0004387	1.80e-04	0.0001815	0.0001233
Wells Fargo	0.0001590	0.0001800	2.24e-04	0.0000734	0.0000605

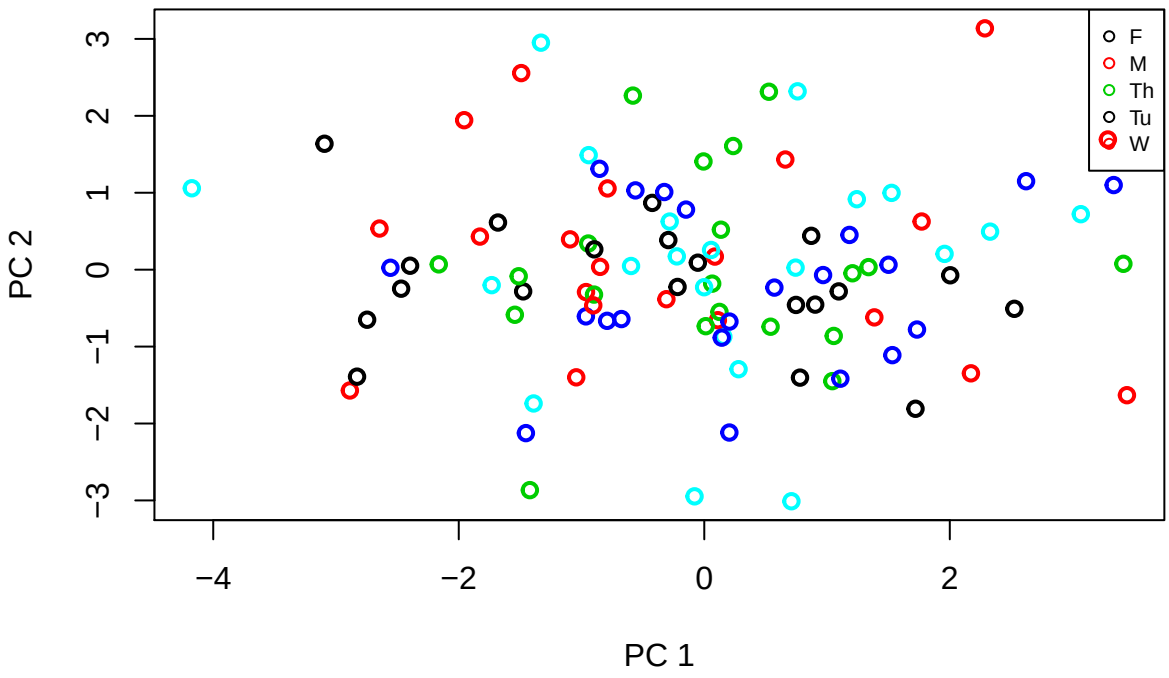
	JP.Morgan	Citibank	Wells.Fargo	Royal.Dutch.Shell	Exxon
Royal Dutch Shell	0.0000641	0.0001815	7.34e-05	0.0007225	0.0005083
Exxon	0.0000890	0.0001233	6.05e-05	0.0005083	0.0007657

Table 2: Table 1: Principal components

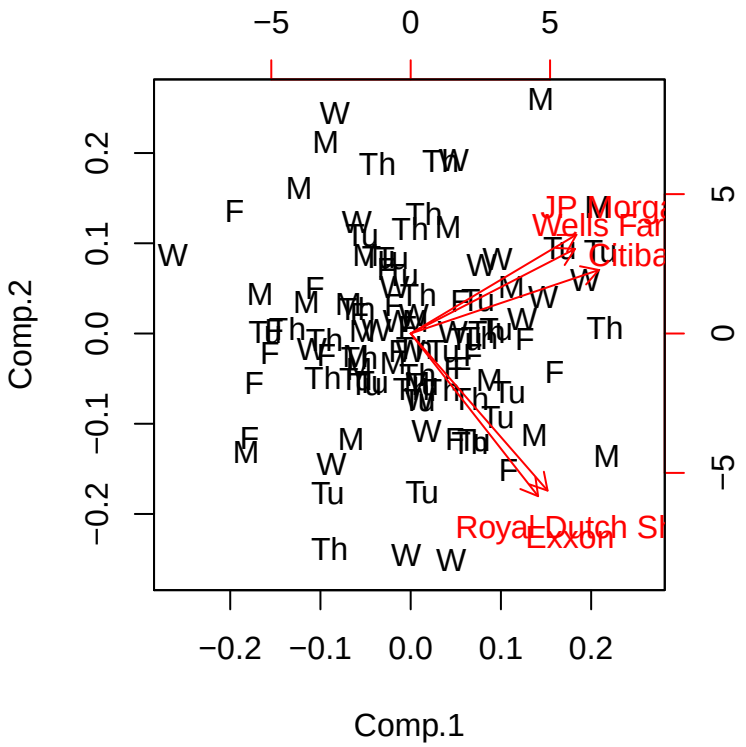
	comp 1	comp 2	comp 3	comp 4	comp 5
Standard deviation	1.5610	1.1800	0.7070	0.632	0.505
Proportion of Variance	0.4875	0.2814	0.1001	0.080	0.051
Cumulative Proportion	0.4875	0.7689	0.8690	0.949	1.000

We comput the scree plot to determin the number of principal components to use. From the plot, we could see that we could use the first 3 components.

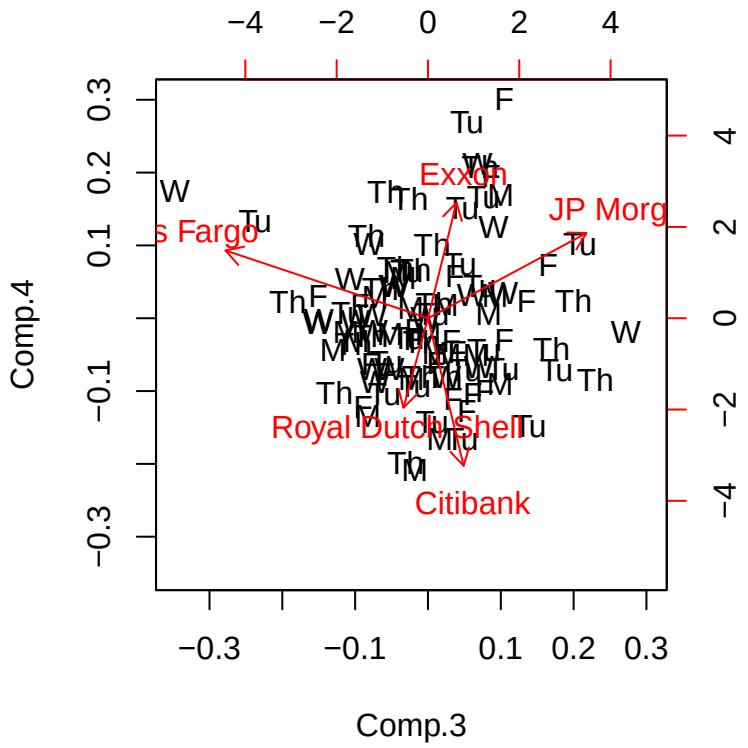




We could see from the plot above that there is no cluster for any particular day.



Looking at figure, we could see that JP Morgan, Wells fargo and Citibank are clustered together towrds Mondays, Tuesdays and Wednesdays while Royal Dutch and Exxon are clustered together in a separate region of Friday.



We dont see much in the plot as the third and fourth principal components do not explain a lot of the variation.

Conclusion

We performed principal component analysis of stock data. We found that the first three principal components capture most of the variance, in fact 87%.

Dataset 3 - Multiple Linear Regression

Introduction

Every year, Forbes, a global business company collects data from publicly listed companies and in turn provide useful indicator (ranking) of which comapnies are the leading public companies. The forbes Global 2000 annual ranking provides ranking index based on four metrics. These metrics are sales, profit, assets and market value. Our aim in this analysis is to formulate a multiple linear regression model based on year 2005 Forbes data in order to predic profits of companies given their asset and sales numbers. The source of the dataset is from Applied Multivariate Statistical Analysis, 5th Edition.

Summary

The dataset consist of the top 10 publicly listed companies according to the Forbes ranking. These companies are Citigroup, General Electric, American Intl Group, Bank of America, HSBC Group, Exxon Mobil, Royal Dutch Shell, BP, ING Group, and Toyota. There are no missing values in the dataset. Our response variable is Profit, while the predictors are assets and sales.

Table 1: Summary Statistics for Forbes Dataset

	Sales	Profits	Assets
Minimum	62.97	8.10	191.1
1st Quartile	92.77	10.96	199.2
Median	130.32	14.94	758.4
Mean	155.60	14.70	710.9
3rd Quartile	239.41	16.93	1090.7
Max	285.06	25.33	1484.1

Figure 1

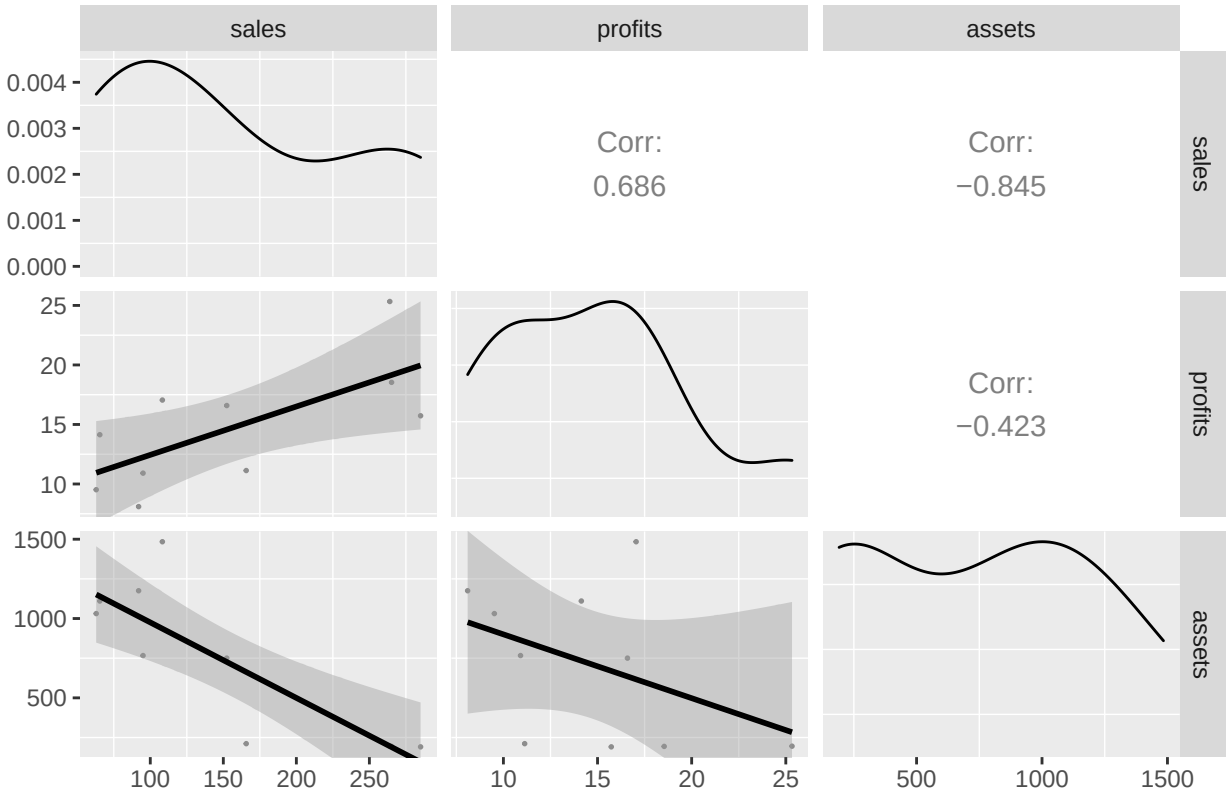


Figure 2

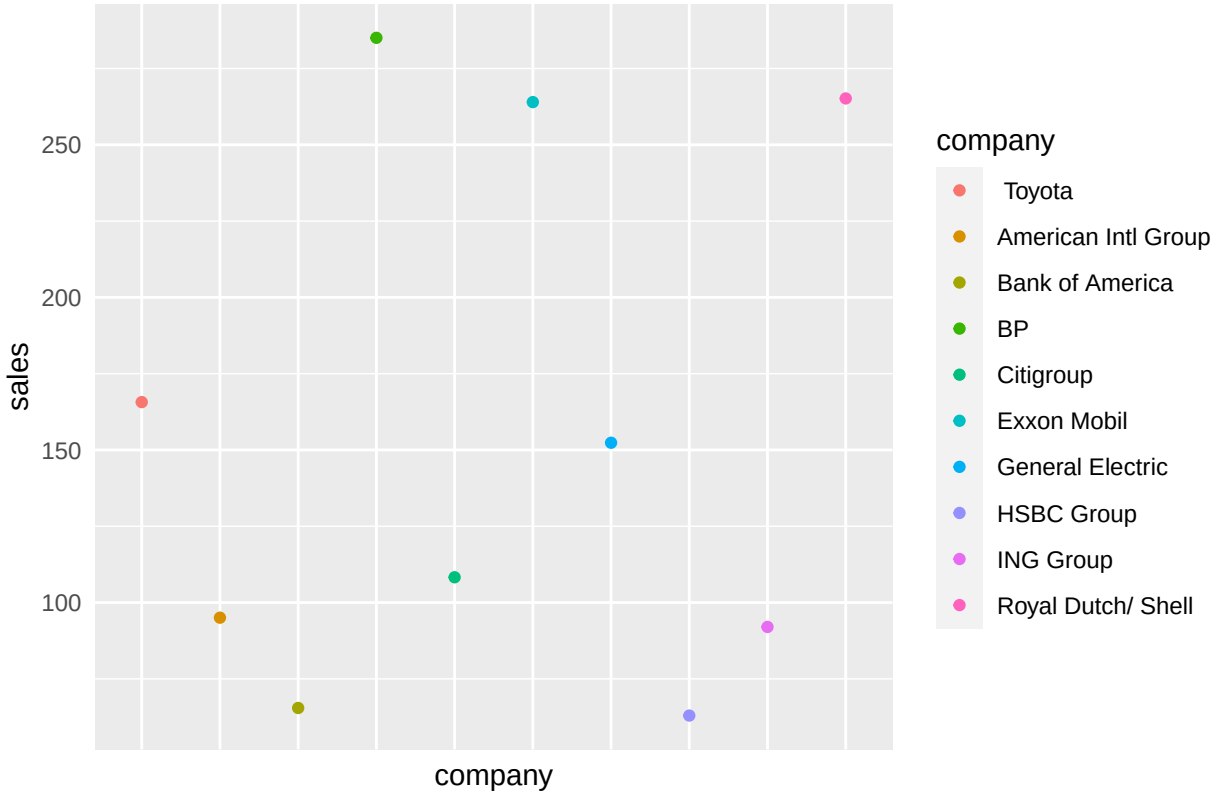
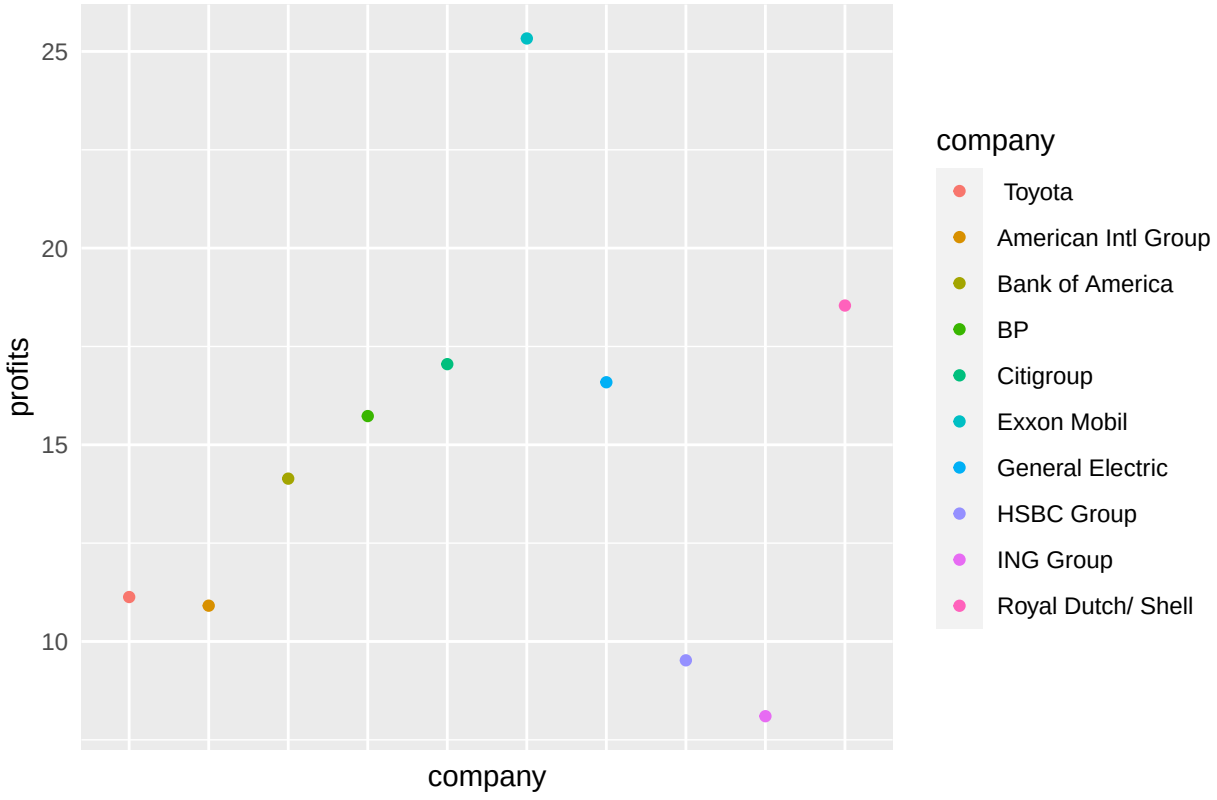


Figure 3



From Figure 1, we observe that the both assets and sales are somewhat strongly correlated. Profit is negatively

correlated with assets and positively correlated with sales. From Figure 2, we observe that BP has the highest sales value and HSBC group has the lowest. Also from Figure 2, we see that Exxon Mobil has the highest profit margin while ING group has the least profit margin.

Analysis

We are interested in predicting profit of ranked companies on the 2005 Forbes annual rankings.

The following model was chosen:

$\mathsf{Y} = \mathsf{X}\beta$

where:

$\mathsf{y_{ij}}$ is profit variable X is design matrix with $\mathsf{X_1}$ as assest and $\mathsf{X_2}$ as Sales

ϵ_{ij} is the error term

The validity of the assumptions for a muitiple linear regression is checked in Figure 4. We could see that the normality assumption is satisfied. For homoskedacity, the number of obsevation is somewhat too small to effectively jusge deviations. So we would assume we can fit a linear regression.

We compute the least squares estimate $\hat{\beta} = (X^t X)X^t Y$

$\hat{\beta} = [0.013, 0.0058, 0.068]^t$

We then compute the Coefficient of determination $R^2 = 0.5568$. This measures of how well the model preforms on obeservations, based on the proportion of total variation explained by the model.

Next, we comput the sample variance $\hat{\sigma^2}$. $\hat{\sigma^2} = 14.92$

By computing $\hat{\sigma^2}(X^t X)^{-1}$, we get the covariance matrix:

$$\mathbf{Cov}(\hat{\beta}) = \begin{bmatrix} 58.391 & 0.036 & -0.20 \\ -0.035 & 0.000025 & 0.00012 \\ -0.20 & 0.00012 & 0.00078 \end{bmatrix}$$

We compute the 95% confidence interval for β_1.

$[\beta_1 - \hat{\sigma^2}\sqrt{\omega_{11}}t_{n-r-1}(\frac{0.05}{2}) + \beta_1 + \hat{\sigma^2}\sqrt{\omega_{11}}t_{n-r-1}(\frac{0.05}{2})] = [-0.00593, 0.01758]$

The 95% confidence intervals for β_j, j= 0,1,2 based on confidence region are

$$[\beta_j - \hat{\sigma^2}\sqrt{\omega_{11}}\sqrt{(r+1)F_{r+1,n-r-1}(0.05))}, \beta_j + \hat{\sigma^2}\sqrt{\omega_{11}}\sqrt{(r+1)F_{r+1,n-r-1}(0.05))}]$$

$\beta_0 \in [-27.5812, 27.60785$
$]$ $\beta_1 \in [-0.0121, 0.0236]$
$\beta_2 \in [-0.03251, 0.1686]$
,

The 95% confidence intervals for β_j, j= 0,1,2 based on Bonferroni correction are

$$[\beta_j - \hat{\sigma^2}\sqrt{\omega_{11}}t_{n-r-1}(\frac{0.05}{2(r+1)})), \beta_j + \hat{\sigma^2}\sqrt{\omega_{11}}t_{n-r-1}(\frac{0.05}{2(r+1)}))]$$

$\beta_0 \in [-23.885, 23.912]$
$\beta_1 \in [-0.0097, 0.0212]$
$\beta_2 \in [-0.019, 0.155]$

We test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$

Let

$$C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The F-statistic is $\frac{1}{\hat{\sigma}^2}\hat{\beta}^t_{(2)}\omega^{-1}_{22}\hat{\beta}_{(2)} = 4.737$

The critical value is $(r-q)F_{r-q,n-r-1}(\alpha) = 2F_{2,7}(0.5) = 9.474$.

Since 4.737 is not greater than 9.474, we do not reject H_0

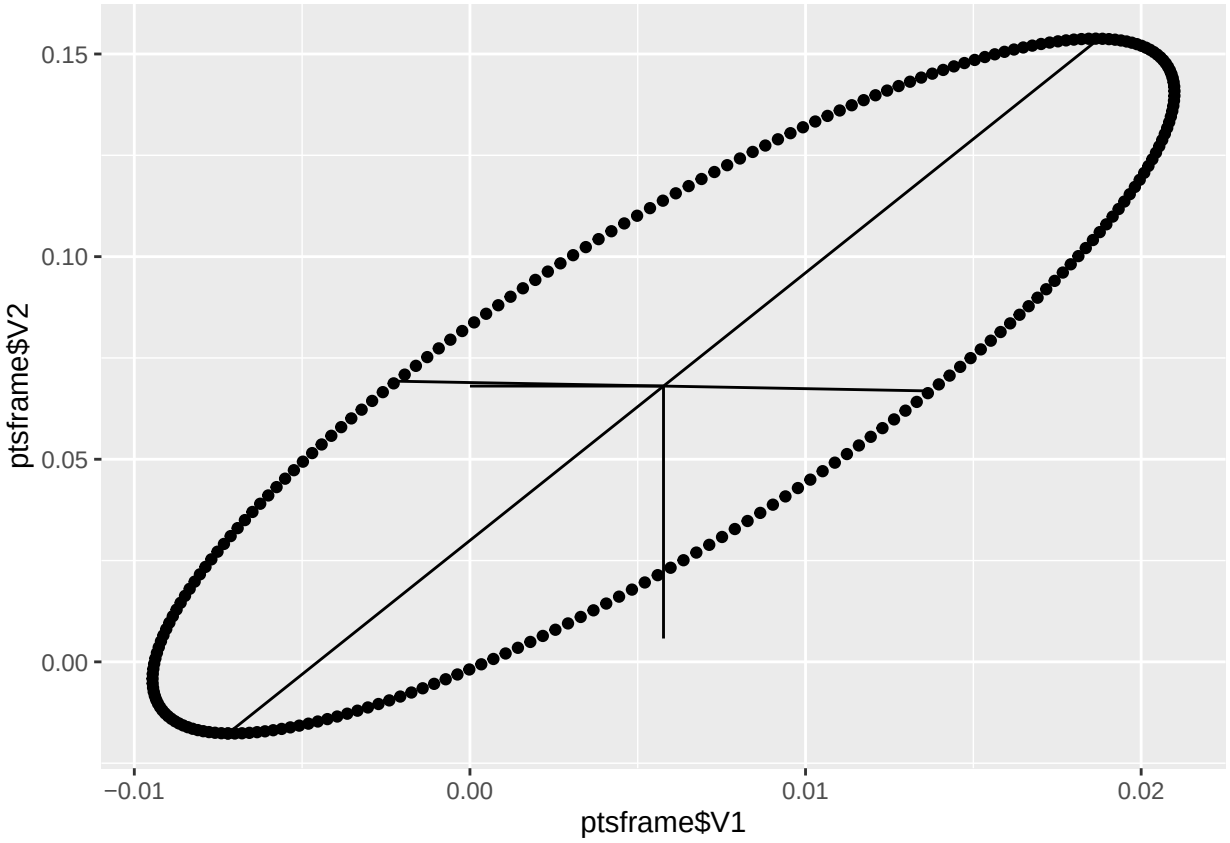
Suppose $x_0 =$

$$C = \begin{bmatrix} 1 \\ 50 \\ 50 \end{bmatrix}$$

The result is $[-10.694, 18.10324]$

The prediction interval for Y_0 given z_0 is given by [-13.34684 , 20.75607]

Below, we plot the confidence region for $[\beta_1, \beta_2]^t$



Conclusion

In analysis, we analyze the top 10 companies in 2005 Forbes annual ranking. We formulated a multiple linear regression model using assets and sales are the predictors and profit as the response variable. We found the least square estimate and found 95% confidence intervals for β_1, 95% confidence intervals for β_1, β_2, and β_3 based on confidence region and Bonferroni correction. We then plotted the confidence region for $[\beta_1, \beta_2]^t$

CODE APPENDIX

1. LDA and Two Sample Test

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(pacman)
p_load(dplyr,table1,ggplot2, GGally, MASS, kableExtra, grid, gridExtra, klaR)
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
bankruptcy <- read.table("T11-4.DAT")
colnames(bankruptcy) <- c("x1","x2", "x3", "x4","population")
label(bankruptcy$x1) = "CF/TD"
label(bankruptcy$x2) = "NI/TA"
label(bankruptcy$x3) <- "CA/CL"
label(bankruptcy$x4) <- "CA/NS"
label(bankruptcy$population) <- "Population"

attach(bankruptcy)
bankruptcy$population = factor(bankruptcy$population)
bankruptcy$x1 = as.numeric(bankruptcy$x1)
bankruptcy$x2 = as.numeric(bankruptcy$x2)
bankruptcy$x3 = as.numeric(bankruptcy$x3)
bankruptcy$x4 = as.numeric(bankruptcy$x4)
## Plot (x1,x2)

ggpairs(bankruptcy[,1:4], aes(color = factor(population)), lower = list(continuous = wrap("smooth", alpha = 0.4, size = 0.3), discrete = "blank", combo="blank"), diag = list(discrete="barDiag", continuous = wrap("densityDiag", alpha=0.5 )), upper = list(combo = wrap("box_no_facet", alpha=0.5), continuous = wrap("cor", size=4, alignPercent=0.8))) + theme(panel.grid.major = element_blank()) + ggtitle("Summary plot for Bankruptcy data")

g2 <- ggplot(bankruptcy, aes(x = x1, y = x2, color = population)) + geom_point() +
  stat_ellipse(aes(x=x1, y= x2, color= x1),type = "norm") +
  theme(legend.position='none') +scale_color_manual(values = c("#00AFBB", "#E7B800"))
# ## Plot (x1,x3)
```

```
g3 <- ggplot(bankruptcy, aes(x = x1, y = x3, color = population)) + geom_point() +
  stat_ellipse(aes(x=x1, y= x3, color= x1),type = "norm") +
  theme(legend.position='none') +scale_color_manual(values = c("#00AFBB", "#E7B800"))
# ## Plot (x1,x4)
g4 <- ggplot(bankruptcy, aes(x = x1, y = x4, color = population)) + geom_point() +
  stat_ellipse(aes(x=x1, y= x4, color= x1),type = "norm") +
  theme(legend.position='none') +scale_color_manual(values = c("#00AFBB", "#E7B800"))
#
grid.arrange(g2,g3,g4, nrow = 2, ncol=2)
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE, include=FALSE}
lda.obj <- lda(population ~ x1+ x2,data=bankruptcy,prior=c(1,1)/2)

plda <- predict(object=lda.obj,newdata=bankruptcy)

# # Confusion matrix
table(population,plda$class)
#
# #plot the decision line
gmean <- lda.obj$prior %*% lda.obj$means

const <- as.numeric(gmean %*%lda.obj$scaling)

slope <- - lda.obj$scaling[1] / lda.obj$scaling[2]

intercept <- const / lda.obj$scaling[2]
#

par(mfrow = c(2,1))
# #Plot decision boundary
plot(bankruptcy[,1:2],pch=rep(c(18,20),each=50),col=rep(c(2,4),each=50))
abline(intercept, slope)
#legend("topright",legend=c("Alaskan","Canadian"),pch=c(18,20),col=c(2,4))
partimat(population~.,data = bankruptcy,method="lda",  main = "LDA Partition Plot")
```

```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
x1_col<- c(-0.069,0.235,3.271)
x2_col <- c(-0.081,0.055,3.367)
ldamat = data.frame(matrix(cbind(x1_col, x2_col), nrow = 3, ncol = 2))
colnames(ldamat) = c("x1","x2")
rownames(ldamat) = c("Group 1 mean", "Group 2 mean","LDA coefficients")
kable(ldamat, digits = 4, format = "pandoc", caption = "Table 1: LDA for x1 and x2")
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
x1_col<- c(-0.069,0.235,2.664)
x3_col <- c(1.3661,2.5936,0.8156)
ldamat = data.frame(matrix(cbind(x1_col, x2_col), nrow = 3, ncol = 2))
colnames(ldamat) = c("x1","x3")
rownames(ldamat) = c("Group 1 mean", "Group 2 mean","LDA coefficients")
kable(ldamat, digits = 4, format = "pandoc", caption = "Table 2: LDA for x1 and x3 ")
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
x1_col<- c(-0.069,0.235,4.6773)
x4_col <- c(0.4376,0.4268,0.01965)
ldamat = data.frame(matrix(cbind(x1_col, x2_col), nrow = 3, ncol = 2))
colnames(ldamat) = c("x1","x4")
rownames(ldamat) = c("Group 1 mean", "Group 2 mean","LDA coefficients")
kable(ldamat, digits = 4, format = "pandoc", caption = "Table 3: LDA for x1 and x4")
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
x2_col <- c(-0.081,0.055,5.496)
x3_col <- c(1.3661,2.5936,0.8896)
ldamat = data.frame(matrix(cbind(x1_col, x2_col), nrow = 3, ncol = 2))
colnames(ldamat) = c("x1","x4")
rownames(ldamat) = c("Group 1 mean", "Group 2 mean","LDA coefficients")
kable(ldamat, digits = 4, format = "pandoc", caption = "Table 4: LDA for x2 and x3")
```
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
x2_col <- c(-0.081,0.055,9.62999)
x4_col <- c(0.4376,0.4268,-0.6980)
ldamat = data.frame(matrix(cbind(x1_col, x2_col), nrow = 3, ncol = 2))
colnames(ldamat) = c("x2","x4")
rownames(ldamat) = c("Group 1 mean", "Group 2 mean","LDA coefficients")
kable(ldamat, digits = 4, format = "pandoc", caption = "Table 5: LDA for x2 and x4")
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
partimat(population~.,data = bankruptcy,method="lda",  main = "LDA Partition Plot")
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
x1_col<- c(-0.069,0.235,0.66124)
x2_col <- c(-0.081,0.055,4.3935)
x3_col <- c(1.3661,2.5936,0.887250)
x4_col <- c(0.4376,0.4268,-1.178500)
ldamat = data.frame(matrix(cbind(x1_col,x2_col,x3_col, x4_col), nrow = 3, ncol = 4))
colnames(ldamat) = c("x1","x2","x3","x4")
rownames(ldamat) = c("Group 1 mean", "Group 2 mean","LDA coefficients")
kable(ldamat, digits = 4, format = "pandoc", caption = "Table 6: LDA for x1, x2, x3 and x4")
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}

## Two sample

bankrupt<- bankruptcy[bankruptcy$population == "0",-c(1,2,5)]
nonbankrupt <-bankruptcy[bankruptcy$population == "1",-c(1,2,5)]

# bankrupt <- iris[iris$Species == "setosa",-c(3,4,5)]
# nonbankrupt <- iris[iris$Species == "versicolor",-c(3,4,5)]
```

```
# now we perform the two-sample Hotelling T^2-test
n<-c(dim(bankrupt)[1],dim(nonbankrupt)[1])
p<- dim(bankruptcy)[2] - 1
xmean1<-colMeans(bankrupt)
xmean2<-colMeans(nonbankrupt)
d<-xmean1-xmean2
S1<-var(bankrupt)
S2<-var(nonbankrupt)
Sp<-((n[1]-1)*S1+(n[2]-1)*S2)/(sum(n)-2)
t2 <- t(d)%*%solve(sum(1/n)*Sp)%*%d


alpha<-0.05
cval <- (sum(n)-2)*p/(sum(n)-p-1)*qf(1-alpha,p,sum(n)-p-1)

```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
# Confidence Region
es<-eigen(sum(1/n)*Sp)
e1<-es$vec %*% diag(sqrt(es$val))
r1<-sqrt(cval)
theta<-seq(0,2*pi,len=250)
v1<-cbind(r1*cos(theta), r1*sin(theta))
pts<-t(d-(e1%*%t(v1)))
plot(pts,type="l",main="Confidence Region for Bivariate Normal",xlab="CF/TD",ylab="NI/TA",asp=1)
segments(0,d[2],d[1],d[2],lty=2) # highlight the center
segments(d[1],0,d[1],d[2],lty=2)

th2<-c(0,pi/2,pi,3*pi/2,2*pi)   #adding the axis
v2<-cbind(r1*cos(th2), r1*sin(th2))
pts2<-t(d-(e1%*%t(v2)))
segments(pts2[3,1],pts2[3,2],pts2[1,1],pts2[1,2],lty=3)
segments(pts2[2,1],pts2[2,2],pts2[4,1],pts2[4,2],lty=3)
```

```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}

# since we reject the null, we use the simultaneous confidence intervals

# to check the significant components


# simultaneous confidence intervals

wd<-sqrt(cval*diag(Sp)*sum(1/n))

Cis<-cbind(d-wd,d+wd)

#Bonferroni simultaneous confidence intervals

wd.b<- qt(1-alpha/(2*p),n[1]+n[2]-2) *sqrt(diag(Sp)*sum(1/n))

Cis.b<-cbind(d-wd.b,d+wd.b)

# both component-wise simultaneous confidence intervals do not contain 0, so they have significant differences.

```

2. Principal Component Analysis

```{r setup, include=FALSE}

knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)

library(GGally)

library(kableExtra)

```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}

stock <-  read.table("T8-4.DAT")

colnames(stock) <- c("JP Morgan","Citibank", "Wells Fargo","Royal Dutch Shell","Exxon")

ggpairs(stock, lower = list(continuous = wrap("smooth", alpha = 0.4, size = 0.3), discrete = "blank", combo="blank"), diag = list(discrete="barDiag", continuous = wrap("densityDiag", alpha=0.5 )), upper = list(combo = wrap("box_no_facet", alpha=0.5), continuous = wrap("cor", size=4, alignPercent=0.8))) + theme(panel.grid.major = element_blank())

```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}

# Sample covariance matrix

scov <- cov(stock)

```
 scov_frame <- data.frame(scov)

kable(scov_frame, format = "pandoc", caption = "Covariance matrix Forbes Dataset")
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
#Principal components

stock.pca <- princomp(stock, cor=TRUE)

stock.pce <- data.frame(matrix(rbind(c(1.561, 1.18, 0.707, 0.632, 0.505 ), c(0.4874546, 0.2814025, 0.1001025, 0.08000632, 0.05103398), c(0.4874546, 0.7688572, 0.8689597, 0.94896602, 1.00000000)), nrow = 3, ncol = 5))

colnames(stock.pce) = c("comp 1", "comp 2", "comp 3","comp 4","comp 5")

rownames(stock.pce) = c("Standard deviation", "Proportion of Variance", "Cumulative Proportion ")

kable(stock.pce, digits = 4, format = "pandoc", caption = "Table 1: Principal components")
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
#Scree plot to determine number of PC to use

# A scree plot:

plot(1:(length(stock.pca$sdev)),  (stock.pca$sdev)^2, type='b',

    main="Scree Plot", xlab="Number of Components", ylab="Eigenvalue Size")

day = factor(rep(c("M", "Tu","W","Th","F"),21))

day = day[1:dim(stock)[1]]
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
plot(stock.pca$scores[,1], stock.pca$scores[,2],

    xlab="PC 1", ylab="PC 2",  lwd=2, col=day)

legend("topright",legend=levels(day),pch=1,col=1:3,cex=0.7)
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
biplot(stock.pca,xlabs=day)
```

```{r echo=FALSE, message=FALSE, warning= FALSE, result=FALSE}
biplot(stock.pca, choices=3:4,xlabs=day)
```

3. Multiple Linear Regression

```{r setup, include=FALSE}

knitr::opts_chunk$set(echo = TRUE)

library(pacman)

p_load(caret, GGally, kableExtra, dplyr, ellipse, gridExtra, grid)

```

```{r echo=FALSE, message=FALSE, result=FALSE,warning= FALSE}


company <- c("Citigroup", "General Electric", "American Intl Group", "Bank of America", "HSBC Group", "Exxon Mobil", "Royal Dutch/ Shell", "BP", "ING Group", " Toyota")


sales <- c(108.28,152.36,95.04,65.45,62.97,263.99,265.19,285.06,92.01,165.68)


profits <- c(17.05,16.59,10.91,14.14,9.52,25.33,18.54,15.73,8.10,11.13)


assets <- c(1484.10,750.33,766.42,1110.46,1031.29,195.26,193.83,191.11,1175.16,211.15)


forbes <-  as.data.frame(matrix(cbind( sales, profits, assets), nrow = 10, ncol = 3))

colnames(forbes) <- c( "sales", "profits", "assets")

forbes$company <- company

forbes$sales <- as.numeric(forbes$sales)

forbes$profits <- as.numeric(forbes$profits)

forbes$assets <- as.numeric(forbes$assets)

```

```{r echo=FALSE, message=FALSE, result=FALSE,warning= FALSE}

forbes_summary <- summary(forbes[,1:3])

sales_sum <- c(62.97,92.77,130.32,155.60,239.41,

285.06)

profits_sum <- c(8.10,10.96,14.94,14.70,16.93,25.33)

assets_sum <- c(191.1,199.2,758.4,710.9,1090.7,1484.1)

five_forbes_summary <- as.data.frame(cbind(sales_sum,profits_sum,assets_sum ))

colnames(five_forbes_summary) = c("Sales","Profits","Assets")

rownames(five_forbes_summary) = c("Minimum", "1st Quartile", "Median", "Mean", "3rd Quartile", "Max")

kable(five_forbes_summary, format = "pandoc", caption = "Summary Statistics for Forbes Dataset")


layout(matrix(c(1,1, 2, 3), nrow = 2, ncol = 2, byrow = TRUE))
```

```
ggpairs(forbes[,1:3], lower = list(continuous = wrap("smooth", alpha = 0.4, size = 0.3), discrete = "blank", combo="blank"), diag = list(discrete="barDiag", continuous = wrap("densityDiag", alpha=0.5 )), upper = list(combo = wrap("box_no_facet", alpha=0.5), continuous = wrap("cor", size=4, alignPercent=0.8))) + theme(panel.grid.major = element_blank()) + ggtitle("Figure 1")


ggplot(forbes, aes( company, sales, color = company)) + geom_point() + theme( axis.text.x = element_blank(), axis.ticks = element_blank()) +ggtitle("Figure 2")

ggplot(forbes, aes( company, profits, color = company)) + geom_point() + theme( axis.text.x = element_blank(), axis.ticks = element_blank()) + ggtitle("Figure 3")


# lay <- c(1,2)

# grid.arrange(grobs=lapply(list(g2,g3),grobTree), layout_matrix = lay)

```


```{r echo=FALSE,include=FALSE, message=FALSE, result=FALSE,warning= FALSE}

model <- lm(profits~assets+sales)

par(mfrow = c(1,2))

plot(model, which = c(2,1), main = "Figure 4")


Z = matrix(cbind(rep(1, 10), assets,  sales), nrow = 10, ncol = 3)

Y =  matrix(profits, nrow = 10, ncol = 1)

n <- length(Y)

r <- dim(Z)[2]-1

```


```{r echo=FALSE, include=FALSE, message=FALSE, result=FALSE,warning= FALSE}

# least square estimates

beta_hat <- solve(t(Z)%*%Z)%*%t(Z)%*%Y

```

explained by the model.

```{r echo=FALSE, include=FALSE, message=FALSE, result=FALSE}

# R^2 statistic

R_square <- 1 - sum((Y - Z%*%beta_hat)^2)/sum((Y-mean(Y))^2)

```

```{r echo=FALSE,include=FALSE, message=FALSE, result=FALSE,warning= FALSE}

# sigma_hat_square

sigma_hat_square <- sum((Y - Z%*%beta_hat)^2)/(n-r-1)

sigma_hat_square
```

```

```{r echo=FALSE,include=FALSE, message=FALSE, result=FALSE}

# estimated covariance of hat{beta}

sigma_hat_square * solve(t(Z)%*%Z)

```


```{r echo=FALSE,include=FALSE, message=FALSE, result=FALSE,warning= FALSE}

# t-test for single coefficient

# H_0: beta_j = 0, H_a: beta_j != 0


j <- 1

t_stat <- (beta_hat[j+1] - 0)/sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1])

t_stat


alpha <- 0.05

cval_t <- qt(1-alpha/2, n-r-1)

cval_t

```


```{r echo=FALSE,include=FALSE, message=FALSE, result=FALSE,warning= FALSE}

# One-at-a-time confidence interval for beta_j


j <- 1

cat('[',

   beta_hat[j+1] - qt(1-alpha/2, n-r-1)*sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),

   ',',

   beta_hat[j+1] + qt(1-alpha/2, n-r-1)*sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),

   ']')

```


The 95% confidence intervals for β_{j}, j= 0,1,2 based on confidence region are $$[\beta_{j} - \hat{\sigma^2}\sqrt{\omega_{11}}\sqrt{(r+1)F_{r+1,n-r-1}(0.05)}), \beta_{j} + \hat{\sigma^2}\sqrt{\omega_{11}}\sqrt{(r+1)F_{r+1,n-r-1}(0.05)})]$$

$\beta_{0} \in [-27.5812 , 27.60785\\]$

$\beta_{1} \in [-0.0121 , 0.0236]\\$

$\beta_{2} \in [-0.03251 , 0.1686]\\$'
```

```{r echo=FALSE,include=FALSE, message=FALSE, result=FALSE,warning= FALSE}
# confidence region based simultaneous confidence intervals


j <- 0
cat('[',
   beta_hat[j+1] - sqrt((r+1)*qf(1-alpha, r+1, n-r-1))*sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
   ',',
   beta_hat[j+1] + sqrt((r+1)*qf(1-alpha, r+1, n-r-1))*sqrt(sigma_hat_square * solve(t(Z)%*%Z)[j+1,j+1]),
   ']')
```