REGRESSION ANALYSIS LAB - DSM1072

ABDUL RAUF

2023-01-24

Contents

NAME:-*ABDUL RAUF*

ROLL:-22*DSMSA*116

ENRL NO:-*GL*6092

1 LAB 1

1.1 ADVERTISING

1.1.1 Read the Advertising data from excel

```
advrt=read.csv("advertising.csv") #to read the advertising data from excel to R.
head(advrt) #Show the head of the advertising data.
```

```
##   X    TV radio newspaper sales
## 1 1 230.1  37.8      69.2  22.1
## 2 2  44.5  39.3      45.1  10.4
## 3 3  17.2  45.9      69.3   9.3
## 4 4 151.5  41.3      58.5  18.5
## 5 5 180.8  10.8      58.4  12.9
## 6 6   8.7  48.9      75.0   7.2
```

```
tail(advrt) #Show the tail part of the advertising data.
```

```
##       X    TV radio newspaper sales
## 195 195 149.7  35.6       6.0  17.3
## 196 196  38.2   3.7      13.8   7.6
## 197 197  94.2   4.9       8.1   9.7
## 198 198 177.0   9.3       6.4  12.8
## 199 199 283.6  42.0      66.2  25.5
## 200 200 232.1   8.6       8.7  13.4
```

1.1.2 FIT THE MODEL

$Sales = \beta_0 + \beta_1 \times TV + \epsilon$

```
y=advrt$sales  #extracting the data of sales and storing values in y
x=advrt$TV     #extracting the data of TV and storing values in x
```

1.1.2.1 Extraction of vaues

1.1.2.2 Calculating b1 i.e beta1 $\beta_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum((x_i-\bar{x})^2)}$

```r
xbar=mean(x)
xbar
```

```
## [1] 147.0425
```

```r
ybar=mean(y)
ybar
```

```
## [1] 14.0225
```

```r
xdev=(x-xbar)
ydev=(y-ybar)
Sxy=sum(xdev*ydev)
Sxx=sum(xdev^2)
b1=Sxy/Sxx
b1 #coefficient of TV
```

```
## [1] 0.04753664
```

1.1.3 Calculating b0 i.e beta0

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \times \bar{x}$

```r
b0=ybar-(b1*xbar)
b0  #value of intercept
```

```
## [1] 7.032594
```

1.1.4 Fitted line for sales due to advertisement on TV

1.1.4.1 yhat $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x$

```r
yhat=b0+b1*x
length(yhat)
```

```
## [1] 200
```

```r
yhat[1:20]  # estimated sales or observed sales based on advertisement on TV
```

```
##  [1] 17.970775  9.147974  7.850224 14.234395 15.627218
##  [6]  7.446162  9.765950 12.746498  7.441409 16.530414
## [11] 10.174765 17.238710  8.163966 11.667416 16.734822
## [16] 16.321253 10.255578 20.409404 10.322129 14.034741
```

1.1.5 Residuals or errors

$\epsilon = y - \hat{y}$

```
e=y-yhat
e[1:20]      #differnce betwn sales and observed sales(i.e sales after advertising on TV)
```

```
##  [1]  4.1292255  1.2520260  1.4497762  4.2656054
##  [5] -2.7272181 -0.2461623  2.0340496  0.4535023
##  [9] -2.6414087 -5.9304143 -1.5747655  0.1612897
## [13]  1.0360344 -1.9674160  2.2651781  6.0787469
## [17]  2.2444222  3.9905958  0.9778709  0.5652593
```

```
length(e)
```

```
## [1] 200
```

1.1.6 Calculate SSE, MSE & Residual statndard error (RSE)

$SSE = \sum e_i^2$=$SSE = \sum((y_i - \hat{y}_i))^2)$ Sum of square due to residual

$MSE = SSE/(n-2)$,where (n-2) is the degree of freedom as 2 degree of freedom get lost as we estimated from sample.

$RSE = \sqrt{(MSE)}$

```
n=length(y)
SSE=sum(e^2)
SSE
```

```
## [1] 2102.531
```

```
MSE=SSE/(n-2)
MSE
```

```
## [1] 10.61884
```

```
RSE=sqrt(MSE)
RSE
```

```
## [1] 3.258656
```

1.1.7 The standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$

$SE(\hat{\beta}_0) = \sqrt{(MSE(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$

$SE(\hat{\beta}_1) = \sqrt{(\frac{MSE}{S_{xx}}}$

```
SEb0=sqrt(MSE/Sxx)
SEb1=sqrt(MSE*(1/n+(xbar)^2/Sxx))
c(SEb0,SEb1)
```

```
## [1] 0.002690607 0.457842940
```

1.1.8 Confidence Interval of $\hat{\beta}_0$ and $\hat{\beta}_1$

CI of $\beta_0 = \hat{\beta}_0 \pm t_(\alpha/2, n-2) * SE(\hat{\beta}_0))$

CI of $\beta_1 = \hat{\beta}_1 \pm t_(\alpha/2, n-2) * SE(\hat{\beta}_1))$

```
cv=qt(p=0.975,df=198)

llb0=b0-cv*SEb0     #lower limit of CI for b0
ulb0=b0+cv*SEb0     #upper limit of CI for b0
cib0=c(llb0,ulb0)   # CI for b0
cib0
```

```
## [1] 7.027288 7.037899
```

```
llb1=b1-cv*SEb1     #lower limit of CI for b1
ulb1=b1+cv*SEb1    #upper limit of CI for b1
cib1=c(llb1,ulb1)   # CI for b1
cib1
```

```
## [1] -0.8553376  0.9504109
```

CI for β_0 is{7.027288,7.037899} CI for β_1 is{-0.8553376,0.9504109}

1.1.9 Coefficient of Determinaton.

$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$

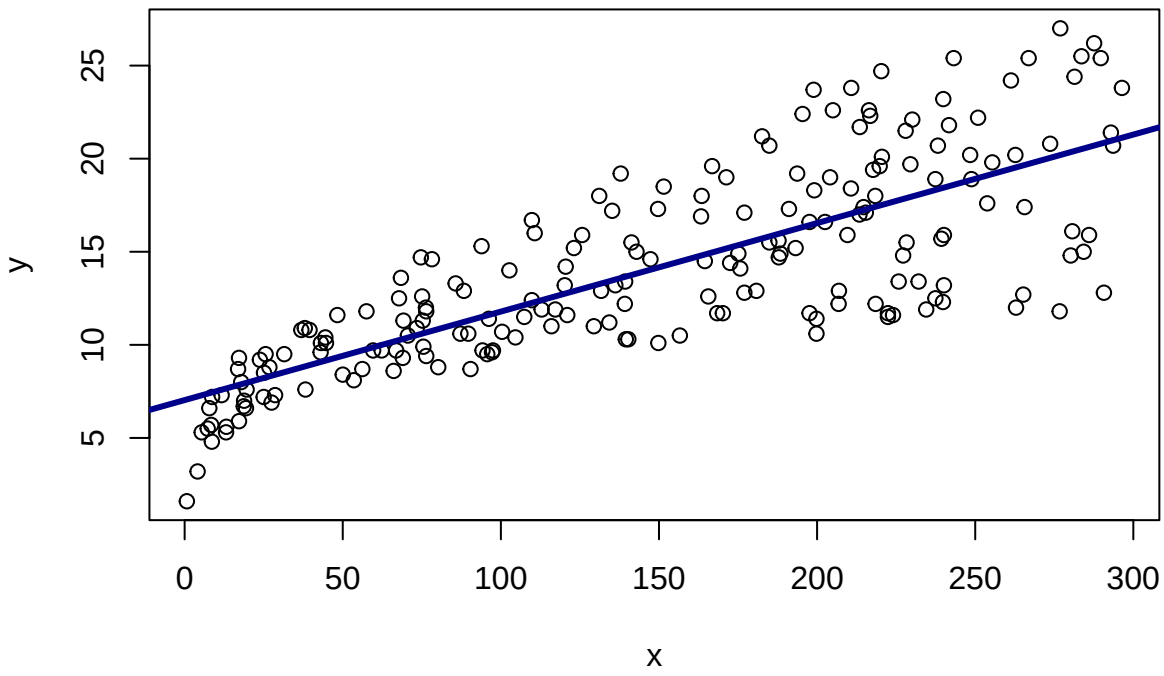
$TSS = \sum(y - \bar{y})^2$

```
TSS=sum((y-ybar)^2)
Rsquared=1-(SSE/TSS)
Rsquared
```

```
## [1] 0.6118751
```

1.1.10 Scatter plot

```
plot(y~x)
abline(7.03259,0.047536,untf = FALSE,lwd=3, col="dark blue")
```



Correlation coefficient

Extent of correlation is explained by correlation coefficient.

```
corl=cor(x,y)
corl
```

```
## [1] 0.7822244
```

1.1.11 Using `lm()` function:

```
m<-lm(y~x,data = advrt)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x, data = advrt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
```

```
## x           0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

1.1.12 CONCLUSION:

- Our model is: $SALES = \beta_0 + \beta_1 \times ADVERTISEMENT\ ON\ TV$

$SALES = 7.03259 + 0.47537 \times ADVERTISEMENT\ ON\ TV$

- we find that form the above calculation and from `lm()` funtion we get the same value for the intercept and coeff. of x or sales

Intercept=7.03259 & **Coeff. of advertisement on TV**=0.47537

- From the estimated value of beta0 and beta1 we can conclude that if the advertisement on TV increases by 1unit then sales get increased by 0.47537 unit.And if we don't advertise on TV then our sales will be 7.03259.
- From the **Scatter plot** we come to the conclusion that sales and advertisement on TV are correlated and they are positively correlated. the distance of the points that are away from the fitted line are the errors.
- As they are correlated so the extent of correlation is explained by **coefficient of correlation** i.e corl=0.7822 , which is nearly equal to 1 that implies that advertisement of TV is highly correlated to sales.
- **p-value** is 2.2e-16 that is very less than 0.05 so advertisement on TV is significant so our null hypothesis is rejected i.e advertisement on TV has no impact on sales but sales get affected by the advertisement on TV .
- **Rsquared** =0.6119 which means 61.19% percent of the variability of sales is explained by advertisement on TV.
- **t value** can be calculated by dividing estimate by standard error (estimate/std.error) t value for x or advertisement on TV is t_value=17.67
- **RSE** Residual standard error tells us that the regression model predicts the sales with the average error of 3.258.
- From the **Confidence interval** we can interpret that we are 95% sure that the correlation between sales and advertisement on TV is between -0.8553376 & 0.9504109

1.2 Assingment-2

1.2.1 FIT THE MODEL

$Sales = \beta_0 + \beta_1 \times NEWSPAPER + \epsilon$

1.2.2 Extraction of vaues

```
y=advrt$sales  #extracting the data of sales and storing values in y
x=advrt$newspaper    #extracting the data of TV and storing values in x
```

1.2.3 Calculating b1 i.e beta1

$\beta_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum((x_i-\bar{x})^2)}$

```
xbar=mean(x)
xbar
```

```
## [1] 30.554
```

```
ybar=mean(y)
ybar
```

```
## [1] 14.0225
```

```
xdev=(x-xbar)
ydev=(y-ybar)
Sxy=sum(xdev*ydev)
Sxx=sum(xdev^2)
b1=Sxy/Sxx
b1 #coefficient of Newspaper
```

```
## [1] 0.0546931
```

1.2.4 Calculating b0 i.e beta0

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \times \bar{x}$

```
b0=ybar-(b1*xbar)
b0  #value of intercept
```

```
## [1] 12.35141
```

1.2.5 Fitted line for sales due to advertisement on Newspaper

1.2.5.1 yhat $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x$

```
yhat=b0+b1*x
yhat[1:20]  # estimated sales or observed sales based on advertisement on newspaper
```

```
##  [1] 16.13617 14.81807 16.14164 15.55095 15.54548 16.45339
##  [7] 13.63669 12.98585 12.40610 13.51090 13.67498 12.57018
## [13] 15.95568 12.74520 14.86729 15.24467 18.58642 15.40328
## [19] 13.35229 13.39605
```

1.2.6 Residuals or errors

$\epsilon = y - \hat{y}$

```
e=y-yhat
e[1:20]  #differnce betwn sales and observed sales(i.e sales after advertising on Newspaper)
```

```
##  [1]  5.963831 -4.418066 -6.841639  2.949047 -2.645484
##  [6] -9.253389 -1.836695  0.214153 -7.606100 -2.910901
## [11] -5.074980  4.829821 -6.755682 -3.045197  4.132710
## [16]  7.155328 -6.086420  8.996718 -2.052291  1.203955
```

1.2.7 Calculate SSE, MSE & Residual statndard error (RSE)

$SSE = \sum e_i^2$=$SSE = \sum((y_i - \hat{y}_i))^2)$ Sum of square due to residual

$MSE = SSE/(n-2)$,where (n-2) is the degree of freedom as 2 degree of freedom get lost as we estimated from sample.

$RSE = \sqrt{(MSE)}$

```
n=length(y)
SSE=sum(e^2)
MSE=SSE/(n-2)
RSE=sqrt(MSE)
RSE
```

```
## [1] 5.09248
```

1.2.8 The standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$

$SE(\hat{\beta}_0) = \sqrt{(MSE(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$

$SE(\hat{\beta}_1) = \sqrt{(\frac{MSE}{S_{xx}}}$

```
SEb0=sqrt(MSE/Sxx)
SEb1=sqrt(MSE*((1/n)+(xbar^2/Sxx)))
c(SEb0,SEb1)
```

```
## [1] 0.01657572 0.62142019
```

1.2.9 Confidence Interval of $\hat{\beta}_0$ and $\hat{\beta}_1$

$\hat{\beta}_0 \pm t_(\alpha/2, n-2) * SE(\hat{\beta}_0))$

$\hat{\beta}_1 \pm t_(\alpha/2, n-2) * SE(\hat{\beta}_1))$

```
cv=qt(p=0.975,df=198)

llb0=b0-cv*SEb0     #lower limit of CI for b0
ulb0=b0+cv*SEb0     #upper limit of CI for b0
cib0=c(llb0,ulb0)   # CI for b0
cib0
```

```
## [1] 12.31872 12.38409
```

```
llb1=b1-cv*SEb1      #lower limit of CI for b1
ulb1=b1+cv*SEb1     #upper limit of CI for b1
cib1=c(llb1,ulb1)    # CI for b1
cib1
```

```
## [1] -1.170758  1.280145
```

CI for β_0 is{7.027288,7.037899} CI for β_1 is{-0.8553376,0.9504109}

1.2.10 Coefficient of Determinaton.

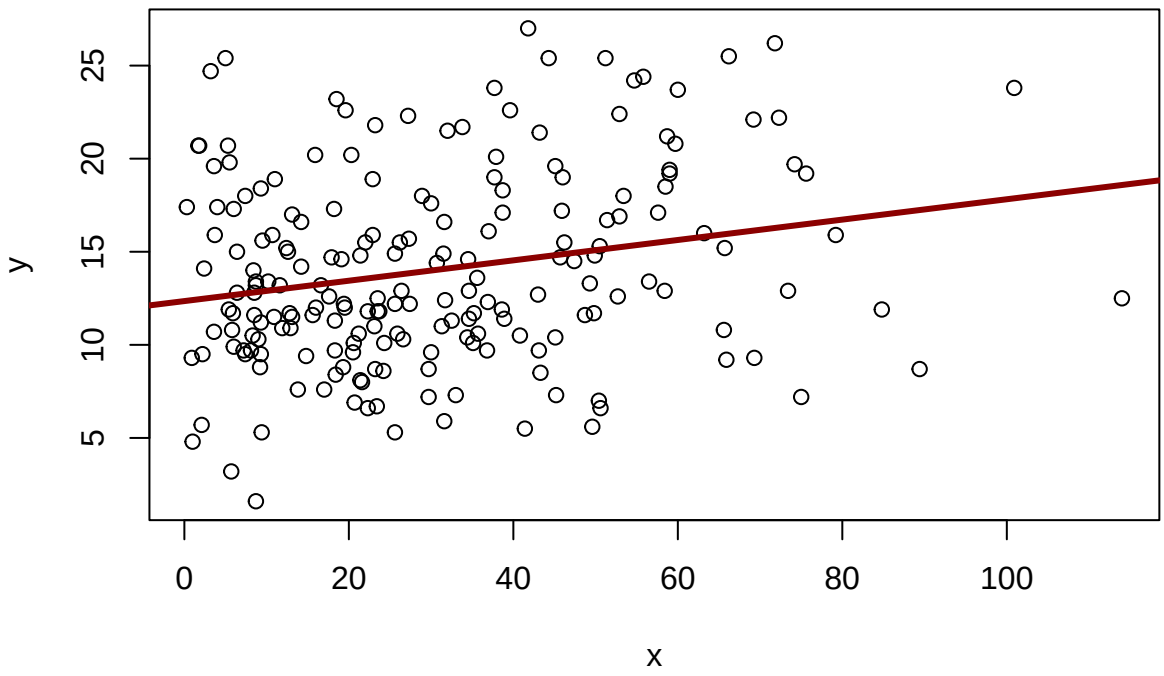
$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$
$TSS = \sum(y - \bar{y})^2$

```
TSS=sum((y-ybar)^2)
Rsquared=1-(SSE/TSS)
Rsquared
```

```
## [1] 0.05212045
```

1.2.11 Scatter plot

```
plot(y~x)
abline(12.35141,0.0546931,untf = FALSE,lwd=3, col="dark red")
```



Correlation coefficient

Extent of correlation is explained by correlation coefficient.

```
corl=cor(x,y)
corl
```

```
## [1] 0.228299
```

1.2.12 Using `lm()` function:

```
m<-lm(y~x,data = advrt)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x, data = advrt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2272  -3.3873  -0.8392   3.5059  12.7751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
```

```
## x            0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.092 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

1.2.13 CONCLUSION:

- Our model is: $SALES = \beta_0 + \beta_1 \times ADVERTISEMENT\ ON\ NEWSPAPER$

$SALES = 12.35141 + 0.0546931 \times ADVERTISEMENT\ ON\ NEWSPAPER$

- we find that form the above calculation and from `lm()` funtion we get the same value for the intercept and coeff. of x or sales

Intercept= 12.35141 & **Coeff. of advertisement on Newspaper**=0.0546931

- From the estimated value of beta0 and beta1 we can conclude that if the advertisement on newspaper increases by 1unit then sales get increased by 0.0546931 unit.And if we don't advertise on newspaper then our sales will remain 12.35141.
- From the **Scatter plot** we come to the conclusion that sales and advertisement on newspaper are correlated and they are positively correlated. the distance of the points that are away from the fitted line are the errors. there are more lines away from the lines and the dots are apart which implies there is low correlation.
- As they are correlated so the extent of correlation is explained by **coefficient of correlation** i.e corl=0.228299, which is nearly equal to 0 that implies that there is low correlation between advertisement of Newspaper and sales.
- **p-value** is 0.001148 that is less than 0.05 so advertisement on Newspaper is significant so our null hypothesis is rejected i.e advertisement on Newspaper has no impact on sales but advertisement on newsaper has some effect on sales but it is low correlation coeff. is near to 0.
- **Rsquared** =0.05212045 which means only 5.21% percent of the variability of sales is explained by advertisement on Newspaper.
- **t value** can be calculated by dividing estimate by standard error (estimate/std.error) t value for x or advertisement on TV is t_value= 3.30
- **RSE** Residual standard error tells us that the regression model predicts the sales with the average error of 5.09248.
- From the **Confidence interval** we can interpret that we are 95% sure that the correlation between sales and advertisement on Newspaper is between -1.170758 & 1.280145

2 LAB 2

2.1 Assingment 1

2.1.1 Install `wooldridge` package and use the data.

We have to install the `wooldridge` package by `install.packages("wooldridge")` then use the package by `require("wooldridge")/library("wooldridge")`

```
require(wooldridge)
```

```
## Loading required package: wooldridge
```

```
head(wage1)
```

```
##   wage educ exper tenure nonwhite female married numdep
## 1 3.10   11     2      0        0      1       0      2
## 2 3.24   12    22      2        0      1       1      3
## 3 3.00   11     2      0        0      0       0      2
## 4 6.00    8    44     28        0      0       1      0
## 5 5.30   12     7      2        0      0       1      1
## 6 8.75   16     9      8        0      0       1      0
##   smsa northcen south west construc ndurman trcommpu
## 1    1        0     0    1        0       0        0
## 2    1        0     0    1        0       0        0
## 3    0        0     0    1        0       0        0
## 4    1        0     0    1        0       0        0
## 5    0        0     0    1        0       0        0
## 6    1        0     0    1        0       0        0
##   trade services profserv profocc clerocc servocc
## 1     0        0        0       0       0       0
## 2     0        1        0       0       0       1
## 3     1        0        0       0       0       0
## 4     0        0        0       0       1       0
## 5     0        0        0       0       0       0
## 6     0        0        1       1       0       0
##      lwage expersq tenursq
## 1 1.131402       4       0
## 2 1.175573     484       4
## 3 1.098612       4       0
## 4 1.791759    1936     784
## 5 1.667707      49       4
## 6 2.169054      81      64
```

```
wage12<-wage1[1:7]
head(wage12)
```

```
##   wage educ exper tenure nonwhite female married
## 1 3.10   11     2      0        0      1       0
## 2 3.24   12    22      2        0      1       1
## 3 3.00   11     2      0        0      0       0
```

```
## 4 6.00    8    44     28        0      0       1
## 5 5.30   12     7      2        0      0       1
## 6 8.75   16     9      8        0      0       1
```

We have extracted first seven variables from **wage1** and then stored the data in **wage12**.

2.1.2 Setting hypothesis

Null Hypothesis: \

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis: \ $H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

2.1.2.1 fit the model.

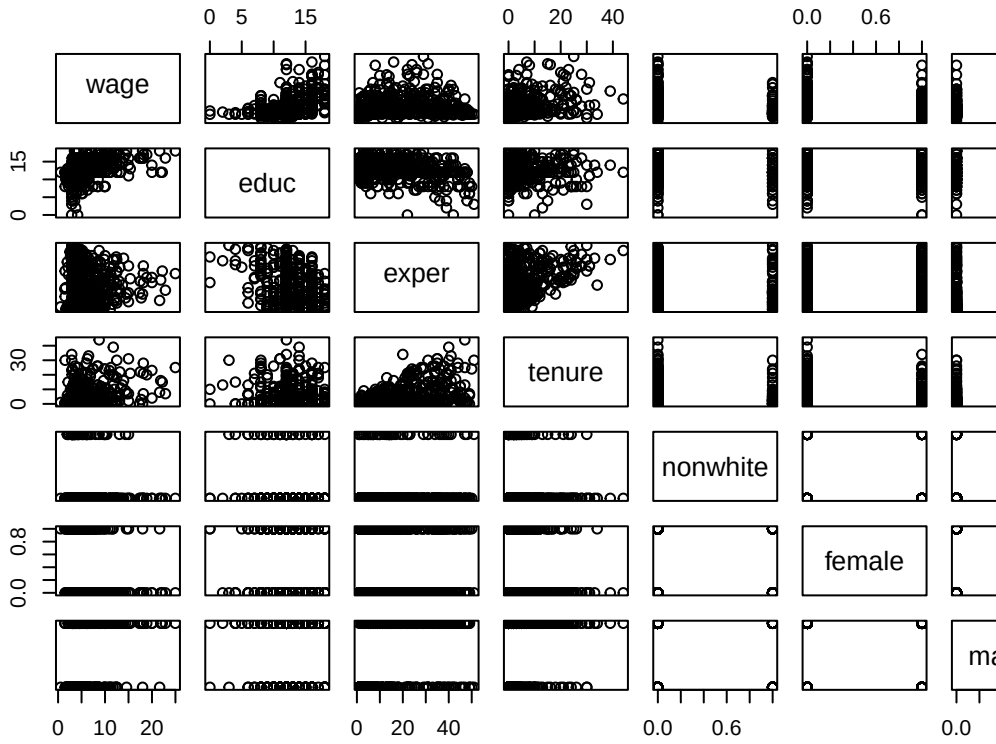
$wage = \beta_0 + \beta_1 \times educ + \beta_2 \times exper + \beta_3 \times tenure + \beta_4 \times nonwhite + \beta_5 \times female + \beta_6 \times married + \epsilon_0$

```
m1<-lm(wage~educ+exper+tenure+nonwhite+female+married , data = wage12)
summary(m1)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + tenure + nonwhite + female +
##     married, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6716 -1.8239 -0.4967  1.0403 13.9209
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.60221    0.73107  -2.192   0.0289 *
## educ         0.55510    0.05006  11.090  < 2e-16 ***
## exper        0.01875    0.01204   1.557   0.1201
## tenure       0.13883    0.02116   6.562 1.29e-10 ***
## nonwhite    -0.06581    0.42657  -0.154   0.8775
## female      -1.74241    0.26682  -6.530 1.57e-10 ***
## married      0.55657    0.28674   1.941   0.0528 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.952 on 519 degrees of freedom
## Multiple R-squared:  0.3682, Adjusted R-squared:  0.3609
## F-statistic: 50.41 on 6 and 519 DF,  p-value: < 2.2e-16
```

- **36.09%** of the variability is explained by this model and this model is significant for 6 & 519 degree of freedom and 5% level of significance but `exper` ,`nonwhite`& `married` are insignificant to this model.

```
pairs(wage12)
```



2.1.2.2 Plotting the graph

From the plot we can see that there is some relationship of wage with `educ`,`expr`,`tenure`.

2.1.3 Names of the variables

```
names(wage12)
```

```
## [1] "wage"     "educ"     "exper"    "tenure"   "nonwhite"
## [6] "female"   "married"
```

2.1.4 Correlation matrix for `wage12` data

```
cor(wage12)
```

```
##                wage        educ       exper      tenure
## wage      1.00000000  0.40590333  0.11290344  0.34688957
## educ      0.40590333  1.00000000 -0.29954184 -0.05617257
## exper     0.11290344 -0.29954184  1.00000000  0.49929145
## tenure    0.34688957 -0.05617257  0.49929145  1.00000000
```

```
## nonwhite -0.03851959 -0.08465433  0.01435563  0.01158880
## female   -0.34009786 -0.08502941 -0.04162597 -0.19791027
## married   0.22881718  0.06888104  0.31698428  0.23988874
##             nonwhite      female     married
## wage     -0.03851959 -0.34009786  0.22881718
## educ     -0.08465433 -0.08502941  0.06888104
## exper     0.01435563 -0.04162597  0.31698428
## tenure    0.01158880 -0.19791027  0.23988874
## nonwhite  1.00000000 -0.01091747 -0.06225929
## female   -0.01091747  1.00000000 -0.16612843
## married  -0.06225929 -0.16612843  1.00000000
```

there is no such high correlation among the variables.we can also check from the VIF

2.1.5 VIF for the given model

```
vif(m1)
```

```
##     educ    exper   tenure nonwhite   female  married
## 1.157103 1.608380 1.407182 1.011547 1.072152 1.182157
```

all the VIF values are not so high which implies that there is no multicolliniarity.

2.1.6 Female as a dummy variable.

$wage = \beta_0 + \beta 1 \times female + \epsilon_0$

as it is a binary variable so for female model will be

$wage = \beta_0 + \beta_1 \times 1$

for male

$wage = \beta_0 + \beta_1 \times 0$
$i.e$
$wage = \beta_0$

```
m2<-lm(wage~female,data = wage12)
summary(m2)
```

```
##
## Call:
## lm(formula = wage ~ female, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5995 -1.8495 -0.9877  1.4260 17.8805
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.0995     0.2100  33.806  < 2e-16 ***
## female       -2.5118     0.3034  -8.279 1.04e-15 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.476 on 524 degrees of freedom
## Multiple R-squared:  0.1157, Adjusted R-squared:  0.114
## F-statistic: 68.54 on 1 and 524 DF,  p-value: 1.042e-15
```

```
#female estimated wage =7.009-2.5118=4.4972
```

- As calculated F-statistic is greater than the tabulated F-statistic at 5% level of significance and 1 & 524 degrees of freedom (i.e 3.9201) so our model is significant.
- The p-value for `female` is less than 0.05 so `female` is significant to the model.
- From above calculation table we come to know that the `wage` of the `female` is estimated to be **4.4972** and the `wage` of the `male` is estimated to be **7.0995**. `wage` of the `male` is higher than that of the `female`.
- **11.4%** of the variability of `wage` is explained by the Model.

2.1.7 Nonwhite as a dummy variable

$wage = \beta_0 + \beta 1 \times nonwhite + \epsilon_0$

as it is a binary variable so for nonwhite model will be

$wage = \beta_0 + \beta_1 \times 1$

for white

$wage = \beta_0 + \beta_1 \times 0$
$i.e$
$wage = \beta_0$

```
m6<-lm(wage~nonwhite,data = wage12)
summary(m6)
```

```
##
## Call:
## lm(formula = wage ~ nonwhite, data = wage12)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.414 -2.526 -1.259  1.026 19.036
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.9442     0.1700  34.961   <2e-16 ***
## nonwhite     -0.4682     0.5306  -0.882    0.378
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.694 on 524 degrees of freedom
## Multiple R-squared:  0.001484,   Adjusted R-squared:  -0.0004218
## F-statistic: 0.7786 on 1 and 524 DF,  p-value: 0.378
```

```
#estimated nonwhite wage =5.9442-0.4682=5.476
```

- As calculated F-statistic is greater than the tabulated F-statistic at 5% level of significance and 1 & 524 degrees of freedom (i.e 3.9201)so our model is insignificant.
- The p-value of nonwhite is greater than 0.05, so nonwhite is insignificant to the model.
- From above calculation table we come to know that the wage of the nonwhite is estimated to be **5.476** and the wage of the white is estimated to be **5.9442** but as it is insignificant so we have to remove nonwhite from our model. wage of the white is higher than that of the nonwhite .
- Negative R squared implies insignificance of explanatory variables i.e `nonwhite` .

2.1.8 Married as a dummy variable

$wage = \beta_0 + \beta 1 \times married + \epsilon_0$

as it is a binary variable so for married model will be

$wage = \beta_0 + \beta_1 \times 1$

for non-married

$wage = \beta_0 + \beta_1 \times 0$
$i.e$
$wage = \beta_0$

```
m7<-lm(wage~married,data = wage12)
summary(m7)
```

```
##
## Call:
## lm(formula = wage ~ married, data = wage12)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.144 -2.181 -1.094  1.406 18.407
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   4.8439     0.2507  19.320  < 2e-16 ***
## married       1.7296     0.3214   5.381 1.12e-07 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.599 on 524 degrees of freedom
## Multiple R-squared:  0.05236,    Adjusted R-squared:  0.05055
## F-statistic: 28.95 on 1 and 524 DF,  p-value: 1.121e-07
```

```
#total estimated married wage = 4.8439+1.7296=6.5735
```

- As calculated F-statistic is greater than the tabulated F-statistic at 5% level of significance and 1 & 524 degrees of freedom (i.e 3.9201) so our model is significant.
- The p-value of married is less than 0.05 so married is significant to the model.
- From above calculation table we come to know that the wage of the married is estimated to be **6.5735** and the wage of the non-married is estimated to be **4.8439**.wage of the non-married is higher than that of the married.
- **5%** of the variability of wage is explained by the model.

2.1.9 forward selection

2.1.9.1 1. we add exper to the model $wage = \beta_0 + \beta_1 \times exper + \epsilon_0$

```
f1<-lm(wage~exper,data = wage12)
summary(f1)
```

```
##
## Call:
## lm(formula = wage ~ exper, data = wage12)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.936 -2.458 -1.112  1.077 18.716
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.37331    0.25699  20.908  < 2e-16 ***
## exper        0.03072    0.01181   2.601  0.00955 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.673 on 524 degrees of freedom
## Multiple R-squared:  0.01275,    Adjusted R-squared:  0.01086
## F-statistic: 6.766 on 1 and 524 DF,  p-value: 0.009555
```

- it is significant to the model as it's p-value is greater than 0.05
- adjusted R-squared is **0.01086** and RSE is **3.673**

2.1.9.2 2. we add `educ` in 1 $wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \epsilon_0$

```
f2<-lm(wage~exper+educ,data = wage12)
summary(f2)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5532 -1.9801 -0.7071  1.2030 15.8370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.39054    0.76657  -4.423 1.18e-05 ***
## exper        0.07010    0.01098   6.385 3.78e-10 ***
## educ         0.64427    0.05381  11.974  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.257 on 523 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.2222
## F-statistic: 75.99 on 2 and 523 DF,  p-value: < 2.2e-16
```

- adjusted r squared get increased so it is significant as it becomes **0.2222** and RSE get reduced **3.257**

2.1.9.3 3. we add `tenure` in 2 $wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \beta_3 \times tenure + \epsilon_0$

```
f3<-lm(wage~exper+educ+tenure,data = wage12)
summary(f3)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ + tenure, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6068 -1.7747 -0.6279  1.1969 14.6536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
## exper        0.02234    0.01206   1.853   0.0645 .
## educ         0.59897    0.05128  11.679  < 2e-16 ***
## tenure       0.16927    0.02164   7.820 2.93e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.084 on 522 degrees of freedom
```

```
## Multiple R-squared:  0.3064, Adjusted R-squared:  0.3024
## F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```

- this model is significant as p-value is less than 0.05 but by the use of the `tenure` in the model `exper` get insignificant so we removed the tenure from our model.

2.1.9.4 4. we add `nonwhite` to our model $wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \beta_4 \times nonwhite + \epsilon_0$

```
f4<-lm(wage~exper+educ+nonwhite,data = wage12)
summary(f4)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ + nonwhite, data = wage12)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.538 -1.982 -0.709  1.205 15.835
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.38683    0.77481  -4.371 1.49e-05 ***
## exper        0.07009    0.01099   6.378 3.95e-10 ***
## educ         0.64412    0.05405  11.917  < 2e-16 ***
## nonwhite    -0.01621    0.47006  -0.034    0.972
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.26 on 522 degrees of freedom
## Multiple R-squared:  0.2252, Adjusted R-squared:  0.2207
## F-statistic: 50.56 on 3 and 522 DF,  p-value: < 2.2e-16
```

- From the above table we find that model is significant as p-value is less than 0.05 .
- But p-value of the `nonwhite` is greater than 0.05 so we remove `nonwhite` from our model.

2.1.9.5 5. we add`married` to our model $wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \beta_5 \times married + \epsilon_0$

```
f5<-lm(wage~exper+educ+married,data = wage12)
summary(f5)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ + married, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7049 -2.0168 -0.5597  1.2077 15.5241
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.37293    0.75990  -4.439 1.11e-05 ***
## exper        0.05688    0.01164   4.888 1.36e-06 ***
## educ         0.61285    0.05423  11.300  < 2e-16 ***
## married      0.98945    0.30920   3.200  0.00146 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.229 on 522 degrees of freedom
## Multiple R-squared:  0.2401, Adjusted R-squared:  0.2357
## F-statistic: 54.97 on 3 and 522 DF,  p-value: < 2.2e-16
```

*From the above table we find that model is significant as p-value is less than 0.05 .

*p-value of the `married` is less than 0.05 so it is significant to our model.

*our adjusted R-squared get increased as it becomes **0.2357** so this model is better than previous model in 2.

2.1.9.6 6. we add `female` to our model $wage = \beta_0 + \beta_1 \times exper + \beta_2 \times educ + \beta_5 \times married + \beta_6 \times female + \epsilon_0$

```
f6<-lm(wage~exper+educ+married+female,data = wage12)
summary(f6)
```

```
##
## Call:
## lm(formula = wage ~ exper + educ + married + female, data = wage12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4057 -1.9042 -0.5982  1.1454 14.6545
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.79066    0.75121  -2.384   0.0175 *
## exper        0.05567    0.01106   5.035 6.59e-07 ***
## educ         0.58332    0.05166  11.292  < 2e-16 ***
## married      0.66024    0.29685   2.224   0.0266 *
## female      -2.06710    0.27221  -7.594 1.45e-13 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.066 on 521 degrees of freedom
## Multiple R-squared:  0.3158, Adjusted R-squared:  0.3105
## F-statistic: 60.12 on 4 and 521 DF,  p-value: < 2.2e-16
```

- From the above table we find that model is significant as p-value is less than 0.05 .
- p-value of the `female` is less than 0.05 so it is significant to our model.
- our adjusted R-squared get increased so this model is better than previous model in 5 i.e **0.3105**.
- RSE is also less than that of model in 5 i.e **3.066** so it is better model.

2.2 Assingment 2

2.2.1 Read the data.

```
credit<-read.csv("credit.csv")
head(credit)
```

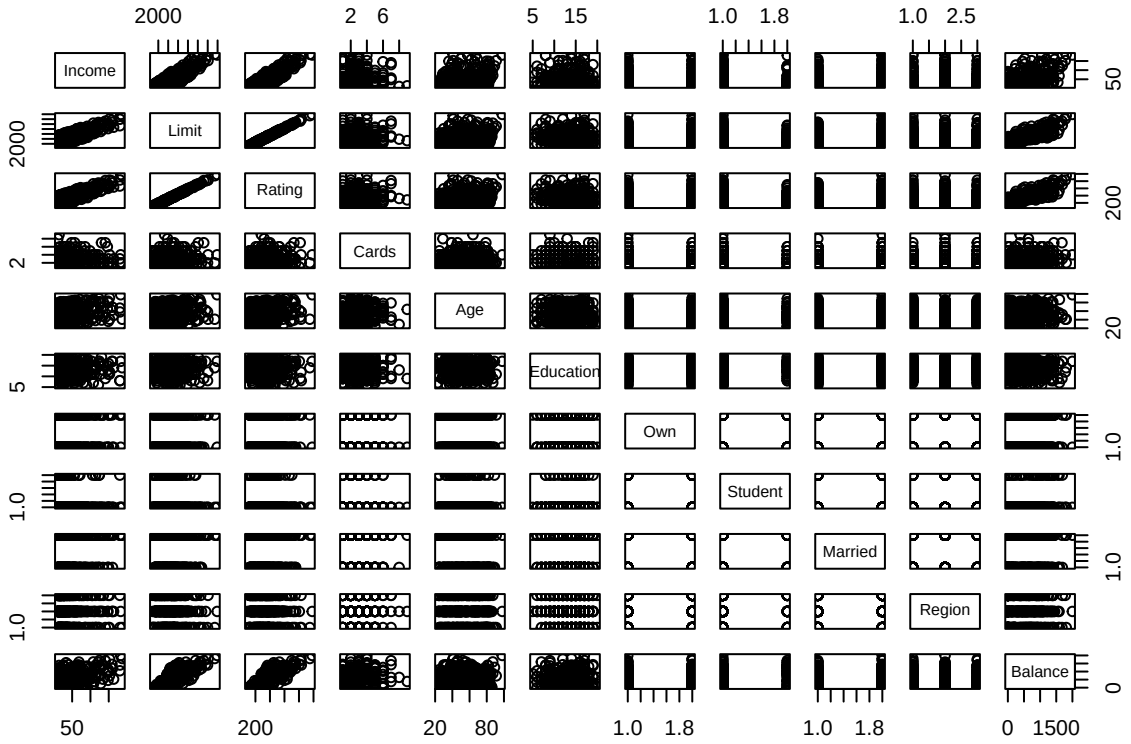
```
##    Income Limit Rating Cards Age Education Own Student
## 1  14.891  3606    283     2  34        11  No      No
## 2 106.025  6645    483     3  82        15 Yes     Yes
## 3 104.593  7075    514     4  71        11  No      No
## 4 148.924  9504    681     3  36        11 Yes      No
## 5  55.882  4897    357     2  68        16  No      No
## 6  80.180  8047    569     4  77        10  No      No
##   Married Region Balance
## 1     Yes  South     333
## 2     Yes   West     903
## 3      No   West     580
## 4      No   West     964
## 5     Yes  South     331
## 6      No  South    1151
```

```
own<-factor(credit$Own,labels=c(0,1)) #transforming the (yes,no) values in (0 & 1) i.e integer
student<-factor(credit$Student,labels=c(0,1)) #transforming the (yes,no) values in (0 & 1) i.e integer
married<-factor(credit$Married,labels=c(0,1)) #transforming the (yes,no) values in (0 & 1) i.e integer
own[1:20]
```

```
##  [1] 0 1 0 1 0 0 1 0 1 1 0 0 1 0 1 1 1 1 1 0
## Levels: 0 1
```

2.2.2 Plot the graph

```
plot(credit)
```



* From the above plot we found that balance has a linear relationship with income,limit,rating and age.

2.2.3 correlation matrix for `credit` data

```
cor(credit[,-7:-10])
```

```
##                Income       Limit      Rating       Cards
## Income     1.00000000  0.79208834  0.79137763 -0.01827261
## Limit      0.79208834  1.00000000  0.99687974  0.01023133
## Rating     0.79137763  0.99687974  1.00000000  0.05323903
## Cards     -0.01827261  0.01023133  0.05323903  1.00000000
## Age        0.17533840  0.10088792  0.10316500  0.04294829
## Education -0.02769198 -0.02354853 -0.03013563 -0.05108422
## Balance    0.46365646  0.86169727  0.86362516  0.08645635
##                   Age    Education      Balance
## Income     0.175338403 -0.027691982  0.463656457
## Limit      0.100887922 -0.023548534  0.861697267
## Rating     0.103164996 -0.030135627  0.863625161
## Cards      0.042948288 -0.051084217  0.086456347
## Age        1.000000000  0.003619285  0.001835119
## Education  0.003619285  1.000000000 -0.008061576
## Balance    0.001835119 -0.008061576  1.000000000
```

- There is high correlation among the variables i.e income & limit,income & rating ,limit & rating.
- As balance is a response variable so there must be a correlation of the variables with them

2.2.4 Setting hypothesis

Null Hypothesis: \

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdot\cdot \quad = \beta_n = 0$

VS

Alternative hypothesis: \ $H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

2.2.4.1 fit the model.

$Balance = \beta_0 + \beta_1 \times income + \beta_2 \times limit + \beta_3 \times rating + \beta_4 \times cards + \beta_5 \times age + \beta_6 \times education + \beta_7 \times balance + \epsilon_0$

```
a2<-lm(Balance~Income+Limit+Rating+Cards+Age+Education+own+student+married, data = credit)
summary(a2)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Education + own + student + married, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -171.66  -75.32  -11.29   54.42  309.98
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -468.40374   34.35512 -13.634  < 2e-16 ***
## Income        -7.80200    0.23395 -33.349  < 2e-16 ***
## Limit          0.19308    0.03268   5.909 7.52e-09 ***
## Rating         1.10227    0.48923   2.253   0.0248 *
## Cards         17.92327    4.33228   4.137 4.31e-05 ***
## Age           -0.63468    0.29325  -2.164   0.0310 *
## Education     -1.11503    1.59592  -0.699   0.4852
## own1         -10.40665    9.90410  -1.051   0.2940
## student1     426.46919   16.67770  25.571  < 2e-16 ***
## married1      -7.01910   10.27803  -0.683   0.4951
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.72 on 390 degrees of freedom
## Multiple R-squared:  0.9549, Adjusted R-squared:  0.9539
## F-statistic: 918.2 on 9 and 390 DF,  p-value: < 2.2e-16
```

- From the above calculated table we came to know that our model is significant as our p-value is less than 0.05
- own and married are insignificant as their p-value is greater than 0.05 except them all the variables are significant.

```
a3<-lm(Balance~Income+Limit+Rating+Cards+Age+student, data = credit)
summary(a2)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Education + own + student + married, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -171.66  -75.32  -11.29   54.42  309.98
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -468.40374   34.35512 -13.634  < 2e-16 ***
## Income        -7.80200    0.23395 -33.349  < 2e-16 ***
## Limit          0.19308    0.03268   5.909 7.52e-09 ***
## Rating         1.10227    0.48923   2.253   0.0248 *
## Cards         17.92327    4.33228   4.137 4.31e-05 ***
## Age           -0.63468    0.29325  -2.164   0.0310 *
## Education     -1.11503    1.59592  -0.699   0.4852
## own1         -10.40665    9.90410  -1.051   0.2940
## student1     426.46919   16.67770  25.571  < 2e-16 ***
## married1      -7.01910   10.27803  -0.683   0.4951
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.72 on 390 degrees of freedom
## Multiple R-squared:  0.9549, Adjusted R-squared:  0.9539
## F-statistic: 918.2 on 9 and 390 DF,  p-value: < 2.2e-16
```

- This model is better than the previous model as in this case adjusted R-squared is greater than the previous one and RSE is minimum in this case.

2.2.5 `student` as a dummy variable

$Balance = \beta_0 + \beta_1 \times student + \epsilon_0$

as it is a binary variable so for student, model will be

$Balance = \beta_0 + \beta_1 \times 1$

for non-student

$Balance = \beta_0 + \beta_1 \times 0$
$i.e$
$Balance = \beta_0$

```
d1<-lm(Balance~student, data = credit)
summary(d1)
```

```
##
## Call:
## lm(formula = Balance ~ student, data = credit)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -876.82 -458.82  -40.87  341.88 1518.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   480.37      23.43   20.50  < 2e-16 ***
## student1      396.46      74.10    5.35 1.49e-07 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 444.6 on 398 degrees of freedom
## Multiple R-squared:  0.06709,    Adjusted R-squared:  0.06475
## F-statistic: 28.62 on 1 and 398 DF,  p-value: 1.488e-07
```

- From the above table we found that our model is significant as our p-value is less than 0.05
- balance of the student is 480.37+396.46 = **876.83** and balance for non-student is **480.37**

2.2.6 `married` as a dummy variable

$Balance = \beta_0 + \beta_1 \times married + \epsilon_0$

```
d2<-lm(Balance~married, data = credit)
summary(d2)
```

```
##
## Call:
## lm(formula = Balance ~ married, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -523.29 -451.03  -60.12  345.06 1481.06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  523.290     36.974  14.153   <2e-16 ***
## married1      -5.347     47.244  -0.113     0.91
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.3 on 398 degrees of freedom
## Multiple R-squared:  3.219e-05,  Adjusted R-squared:  -0.00248
## F-statistic: 0.01281 on 1 and 398 DF,  p-value: 0.9099
```

- From the above calculated table we found that our model is insignificant as our p-value is greater than 0.05.

2.2.7 `own` as a dummy variable

$Balance = \beta_0 + \beta_1 \times own + \epsilon_0$

as it is a binary variable so for own, model will be

$Balance = \beta_0 + \beta_1 \times 1$

for non-own

$Balance = \beta_0 + \beta_1 \times 0$
$i.e$
$Balance = \beta_0$

```
d3<-lm(Balance~own, data = credit)
summary(d3)
```

```
##
## Call:
## lm(formula = Balance ~ own, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -529.54 -455.35  -60.17  334.71 1489.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   509.80      33.13  15.389   <2e-16 ***
## own1           19.73      46.05   0.429    0.669
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.2 on 398 degrees of freedom
## Multiple R-squared:  0.0004611,  Adjusted R-squared:  -0.00205
## F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

- From the above calculated we found that our model is in significant as it is greater than 0.05.

2.3 Assingment 3

2.3.1 read the data from excel to r

```
chemical<-read.csv("ChemicalData.csv")
head(chemical)
```

```
##   acidtemp acidconc watertemp sulfideconc amtofbleach
## 1       35      0.3        82         0.2         0.3
## 2       35      0.3        82         0.3         0.5
## 3       35      0.3        88         0.2         0.5
```

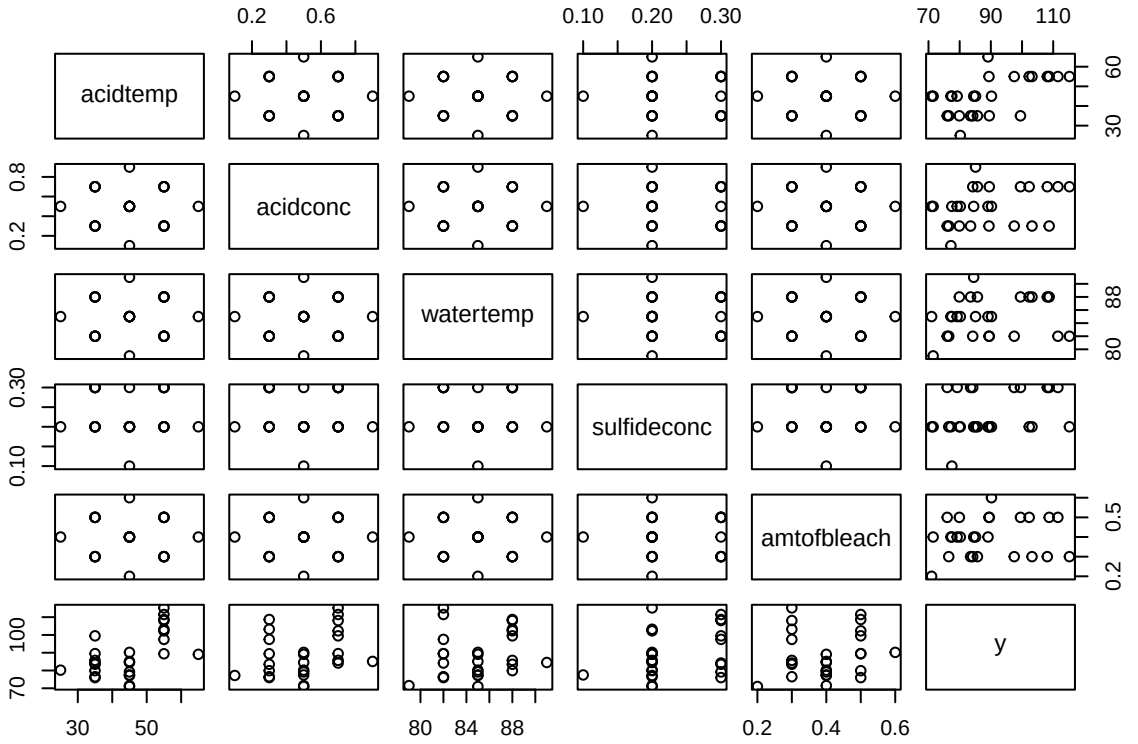
```
## 4       35      0.3        88         0.3         0.3
## 5       35      0.7        82         0.2         0.5
## 6       35      0.7        82         0.3         0.3
##      y
## 1 76.5
## 2 76.0
## 3 79.9
## 4 83.5
## 5 89.5
## 6 84.2
```

```
names(chemical)
```

```
## [1] "acidtemp"    "acidconc"    "watertemp"
## [4] "sulfideconc" "amtofbleach" "y"
```

2.3.2 plot the data

```
plot(chemical)
```



2.3.3 correlation matrix of the chemical data

```
cor(chemical)
```

```
##              acidtemp      acidconc watertemp
## acidtemp    1.0000000  0.000000e+00 0.0000000
## acidconc    0.0000000  1.000000e+00 0.0000000
## watertemp   0.0000000  0.000000e+00 1.0000000
## sulfideconc 0.0000000 -3.491215e-17 0.0000000
## amtofbleach 0.0000000  0.000000e+00 0.0000000
## y           0.5709244  3.098757e-01 0.1822235
##                sulfideconc   amtofbleach         y
## acidtemp     0.000000e+00  0.000000e+00 0.5709244
## acidconc    -3.491215e-17  0.000000e+00 0.3098757
## watertemp    0.000000e+00  0.000000e+00 0.1822235
## sulfideconc  1.000000e+00 -5.079390e-17 0.3303949
## amtofbleach -5.079390e-17  1.000000e+00 0.1318009
## y            3.303949e-01  1.318009e-01 1.0000000
```

There is a correlation between `acidtemp` and `y` i.e **0.5709244**(57% correlation between acidtemp and y) ,`sulfideconc` and `y` i.e **0.3303949** (33% correlation between sulfideconc and y) & `acidtemp` and `y` i.e **0.3098757** (30% correlation between acidconc and y)

2.3.4 Setting hypothesis

Null Hypothesis: \

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis: \ $H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

2.3.4.1 fit the model.

$y = \beta_0 + \beta_1 \times acidtemp + \beta_2 \times acidtemp + \beta_3 \times watertemp + \beta_4 \times sulfideconc + \beta_5 \times amtofbleach + \epsilon_0$

```
chem<-lm(y~.,data = chemical)
summary(chem)
```

```
##
## Call:
## lm(formula = y ~ ., data = chemical)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2133  -5.3674   0.0128   5.1365  21.0837
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -46.6289    55.4735  -0.841 0.410530
## acidtemp      0.7454     0.1888   3.948 0.000795 ***
## acidconc     20.2292     9.4409   2.143 0.044620 *
## watertemp     0.7931     0.6294   1.260 0.222161
```

```
## sulfideconc  76.9694    33.6904   2.285 0.033394 *
## amtofbleach  17.2083    18.8817   0.911 0.372952
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.25 on 20 degrees of freedom
## Multiple R-squared:  0.5817, Adjusted R-squared:  0.4771
## F-statistic: 5.563 on 5 and 20 DF,  p-value: 0.002272
```

- from the above calculated table we can find that the model is significant as p-value is less than 0.05
- p-value of intercept, watertemp,amtofbleach is greater than 0.05 so it is insignificant but rest of the variables acidtemp,acidconc,sulfideconc are significant.

2.3.5 removing amtofbleach

```
chem1<-lm(y~acidtemp+acidconc+watertemp+sulfideconc, data = chemical)
summary(chem1)
```

```
##
## Call:
## lm(formula = y ~ acidtemp + acidconc + watertemp + sulfideconc,
##     data = chemical)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7163  -4.5070   0.8337   5.5026  19.3628
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.7456    54.7349  -0.726 0.475765
## acidtemp      0.7454     0.1881   3.964 0.000708 ***
## acidconc     20.2292     9.4027   2.151 0.043235 *
## watertemp     0.7931     0.6268   1.265 0.219676
## sulfideconc  76.9694    33.5542   2.294 0.032212 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.213 on 21 degrees of freedom
## Multiple R-squared:  0.5643, Adjusted R-squared:  0.4814
## F-statistic: 6.801 on 4 and 21 DF,  p-value: 0.001126
```

- From the above table we find that our model is significant but intercept and watertemp is insignificant to the model as their p-value is greater than 0.05.so we remove watertemp from our model.

2.3.6 removing watertemp

```r
chem2<-lm(y~acidtemp+acidconc+sulfideconc, data = chemical)
summary(chem2)
```

```
##
## Call:
## lm(formula = y ~ acidtemp + acidconc + sulfideconc, data = chemical)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7163  -5.6578   0.9352   5.3610  16.9837
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.6641    12.6974   2.179 0.040346 *
## acidtemp      0.7454     0.1906   3.911 0.000749 ***
## acidconc     20.2292     9.5302   2.123 0.045278 *
## sulfideconc  76.9694    34.0091   2.263 0.033834 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.338 on 22 degrees of freedom
## Multiple R-squared:  0.5311, Adjusted R-squared:  0.4672
## F-statistic: 8.307 on 3 and 22 DF,  p-value: 0.0007028
```

- From the calculated table we find that our model is significant and now our all the variables as well as intercept get significant as p-value is less than 0.05 .
- Now our model is perfect fit.

3 LAB 3

3.1 Assingment 1

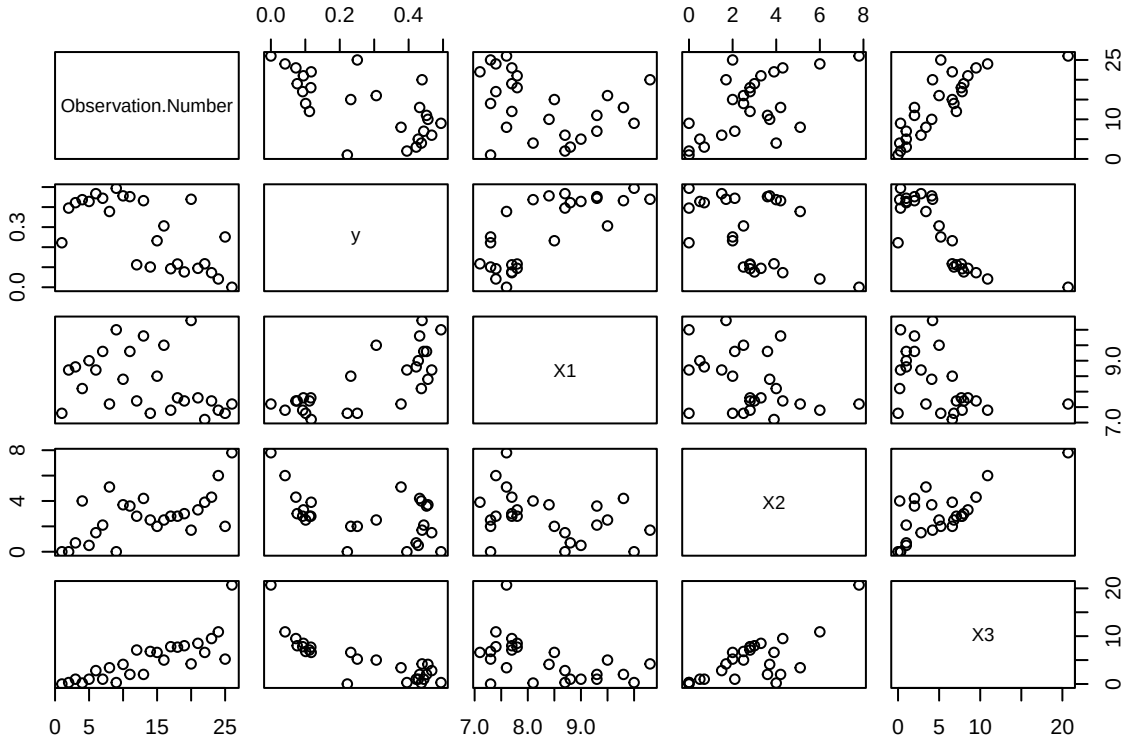
3.1.1 Reading the data from excel.

```r
solubility<-read.csv("solubility.csv")
head(solubility)
```

```
##   Observation.Number     y  X1  X2  X3
## 1                  1 0.222 7.3 0.0 0.0
## 2                  2 0.395 8.7 0.0 0.3
## 3                  3 0.422 8.8 0.7 1.0
## 4                  4 0.437 8.1 4.0 0.2
## 5                  5 0.428 9.0 0.5 1.0
## 6                  6 0.467 8.7 1.5 2.8
```

3.1.2 Plotting the data

```
plot(solubility)
```



- there is no linear relationship between the variables.
- there is a quadratic relationship between the variables.

3.1.3 Model 1

3.1.4 Setting hypothesis

Null Hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis:
$H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

```
sol<- lm(y~X1+X2+X3+I(X1^2),data = solubility)
summary(sol)
```

3.1.4.1 fit the model.

```
##
## Call:
## lm(formula = y ~ X1 + X2 + X3 + I(X1^2), data = solubility)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07613 -0.05257 -0.01167  0.03038  0.10756
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.237128   1.236704  -1.809   0.0848 .
## X1           0.525191   0.289018   1.817   0.0835 .
## X2           0.023573   0.009949   2.369   0.0275 *
## X3          -0.027614   0.004396  -6.282 3.14e-06 ***
## I(X1^2)     -0.025462   0.016751  -1.520   0.1434
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06394 on 21 degrees of freedom
## Multiple R-squared:  0.8799, Adjusted R-squared:  0.8571
## F-statistic: 38.48 on 4 and 21 DF,  p-value: 2.208e-09
```

- our model is significant but intercept,x1 and I(x1^2) is insignificant to the model.
- **85.71%** of variability of y is explained by the model.

3.1.5 Model 2

3.1.5.1 Setting hypothesis

Null Hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis:
$H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

```
sol3<- lm(y~X1+X2+X3+I(X1^2)+I(X2^2)+I(X3^2),data = solubility)
summary(sol3)
```

3.1.5.2 fit the model.

```
##
## Call:
## lm(formula = y ~ X1 + X2 + X3 + I(X1^2) + I(X2^2) + I(X3^2),
##     data = solubility)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.072137 -0.036041 -0.008668  0.007245  0.118944
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.8446143  1.2420061  -1.485  0.15390
## X1           0.4491296  0.2889529   1.554  0.13660
## X2           0.0155739  0.0328394   0.474  0.64073
## X3          -0.0385807  0.0119318  -3.233  0.00437 **
## I(X1^2)     -0.0214860  0.0166719  -1.289  0.21296
## I(X2^2)      0.0011441  0.0057537   0.199  0.84450
## I(X3^2)      0.0005904  0.0008048   0.734  0.47211
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06089 on 19 degrees of freedom
## Multiple R-squared:  0.9015, Adjusted R-squared:  0.8704
## F-statistic: 28.97 on 6 and 19 DF,  p-value: 1.376e-08
```

- From the above calculated table we find that our model is significant as the p value of the model is less than **0.05** so our null hyothesis get rejected that other variables have no impact on the model.
- **87.04%** of variability of y is explained by the model.
- *t-statistic* for the parameter is

• for intercept $= \beta_0/SE(\beta_0)$ which is equal to **-1.485** and from the p-value we find that it is insignificant to our model.

• for X_1 $= \beta_1/SE(\beta_1)$ which is equal to **1.554** and from the p-value we find that it is insignificant to our model.

• for X_2 $= \beta_2/SE(\beta_2)$ which is equal to **0.474** and from the p-value we find that it is insignificant to our model.

• for X_3 $= \beta_3/SE(\beta_3)$ which is equal to **-3.233** and from the p-value we find that it is significant to our model.

• for $I(X_1^2)$ $= \beta_4/SE(\beta_4)$ which is equal to **-1.289** and from the p-value we find that it is insignificant to our model.

• for $I(X_2^2)$ $= \beta_5/SE(\beta_5)$ which is equal to **0.199** and from the p-value we find that it is insignificant to our model.

• for $I(X_3^2)$ $= \beta_6/SE(\beta_6)$ which is equal to **0.734** and from the p-value we find that it is insignificant to our model.

- All the pameters except X_3 are insignificant so our model is not good fit or best fit.

3.1.6 Model 3

3.1.6.1 Setting hypothesis Null Hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis:
H_1 : *At least one of the* β_i *'s are not zero*

```
sol3<- lm(y~X1+X3+I(X1^2)+I(X2^2)+I(X3^2),data = solubility)
summary(sol3)
```

3.1.6.2 fit the model.

```
##
## Call:
## lm(formula = y ~ X1 + X3 + I(X1^2) + I(X2^2) + I(X3^2), data = solubility)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.070515 -0.038575 -0.006143  0.008974  0.119647
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.9953841  1.1771337  -1.695 0.105567
## X1           0.4849071  0.2734725   1.773 0.091435 .
## X3          -0.0340906  0.0071194  -4.788 0.000112 ***
## I(X1^2)     -0.0234927  0.0158105  -1.486 0.152902
## I(X2^2)      0.0037557  0.0016344   2.298 0.032489 *
## I(X3^2)      0.0002686  0.0004241   0.633 0.533715
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0597 on 20 degrees of freedom
## Multiple R-squared:  0.9003, Adjusted R-squared:  0.8754
## F-statistic: 36.12 on 5 and 20 DF,  p-value: 2.385e-09
```

- From the above calculated table we find that our model is significant as the p value of the model is less than **0.05** so our null hyothesis get rejected that other variables have no impact on the model.
- **87.54%** of variability ofy is explained by the model.
- *t-statistic* for the parameter is

• for intercept $= \beta_0/SE(\beta_0)$ which is equal to **-1.695** and from the p-value we find that it is insignificant to our model.

• for X_1 $= \beta_1/SE(\beta_1)$ which is equal to **1.773** and from the p-value we find that it is insignificant to our model.

• for X_3 $= \beta_3/SE(\beta_3)$ which is equal to **-4.788** and from the p-value we find that it is significant to our model.

• for $I(X_1^2)$ $= \beta_4/SE(\beta_4)$ which is equal to **-1.486** and from the p-value we find that it is insignificant to our model.

• for $I(X_2^2) = \beta_5/SE(\beta_5)$ which is equal to **2.298** and from the p-value we find that it is significant to our model.

• for $I(X_3^2) = \beta_6/SE(\beta_6)$ which is equal to **0.633** and from the p-value we find that it is insignificant to our model.

- it has the highest Adj R-sq and lowest RSE but most of the parameters are insignificant to the model.

3.1.7 Model 4

3.1.7.1 Setting hypothesis

Null Hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis:
$H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

```
sol4<- lm(y~X1+X3+I(X2^2),data = solubility)
summary(sol4)
```

3.1.7.2 fit the model.

```
##
## Call:
## lm(formula = y ~ X1 + X3 + I(X2^2), data = solubility)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08962 -0.03076 -0.00990  0.02206  0.13375
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.303539   0.132433  -2.292  0.03184 *
## X1           0.082623   0.014849   5.564 1.36e-05 ***
## X3          -0.031493   0.004401  -7.156 3.57e-07 ***
## I(X2^2)      0.004447   0.001388   3.205  0.00408 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06097 on 22 degrees of freedom
## Multiple R-squared:  0.8856, Adjusted R-squared:   0.87
## F-statistic: 56.78 on 3 and 22 DF,  p-value: 1.605e-10
```

- From the above calculated table we find that our model is significant as the p value of the model is less than **0.05** so our null hyothesis get rejected that other variables have no impact on the model.
- **87%** of variability of y is explained by the model.

- *t-statistic* for the parameter is

• for intercept $= \beta_0/SE(\beta_0)$ which is equal to **-2.292** and from the p-value we find that it is significant to our model.

• for X_1 $= \beta_1/SE(\beta_1)$ which is equal to **5.564** and from the p-value we find that it is significant to our model.

• for X_3 $= \beta_3/SE(\beta_3)$ which is equal to **-7.156** and from the p-value we find that it is significant to our model.

• for $I(X_2^2)$ $= \beta_5/SE(\beta_5)$ which is equal to **3.205** and from the p-value we find that it is significant to our model.

- In this model all the parameters are significant and it has nearly the same Adj R-sq value as that of the highest Adj R-sq and lowest RSE and this is good as respect to the other models.

3.1.8 Model 5

3.1.8.1 Setting hypothesis

Null Hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis:
$H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

```
sol5<- lm(y~X1+X2+X3+I(X1^2+X2^2+X3^2),data = solubility)
summary(sol5)
```

3.1.8.2 fit the model.

```
##
## Call:
## lm(formula = y ~ X1 + X2 + X3 + I(X1^2 + X2^2 + X3^2), data = solubility)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08941 -0.02975 -0.01368  0.02105  0.13827
##
## Coefficients:
##                          Estimate Std. Error t value
## (Intercept)           -0.2078136  0.1512156  -1.374
## X1                     0.0659969  0.0174256   3.787
## X2                     0.0185572  0.0097907   1.895
## X3                    -0.0410966  0.0070060  -5.866
## I(X1^2 + X2^2 + X3^2)  0.0007026  0.0003177   2.211
##                         Pr(>|t|)
## (Intercept)             0.18384
## X1                      0.00108 **
```

```
## X2                     0.07189 .
## X3                    8.02e-06 ***
## I(X1^2 + X2^2 + X3^2)  0.03823 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06067 on 21 degrees of freedom
## Multiple R-squared:  0.8919, Adjusted R-squared:  0.8713
## F-statistic: 43.32 on 4 and 21 DF,  p-value: 7.422e-10
```

- This model is also significant but two of the parameters are insignificant to our model.
- **87.13%** of variability ofy is explained by the model.

3.1.9 Correlation of the variables.

```
cor(solubility)
```

```
##                     Observation.Number          y
## Observation.Number          1.0000000 -0.6909653
## y                          -0.6909653  1.0000000
## X1                         -0.3299570  0.7620911
## X2                          0.5933689 -0.4670590
## X3                          0.8037425 -0.8117565
##                            X1         X2         X3
## Observation.Number -0.3299570  0.5933689  0.8037425
## y                   0.7620911 -0.4670590 -0.8117565
## X1                  1.0000000 -0.3717451 -0.4925686
## X2                 -0.3717451  1.0000000  0.7243160
## X3                 -0.4925686  0.7243160  1.0000000
```

```
cor(solubility[c(-1,-4)])
```

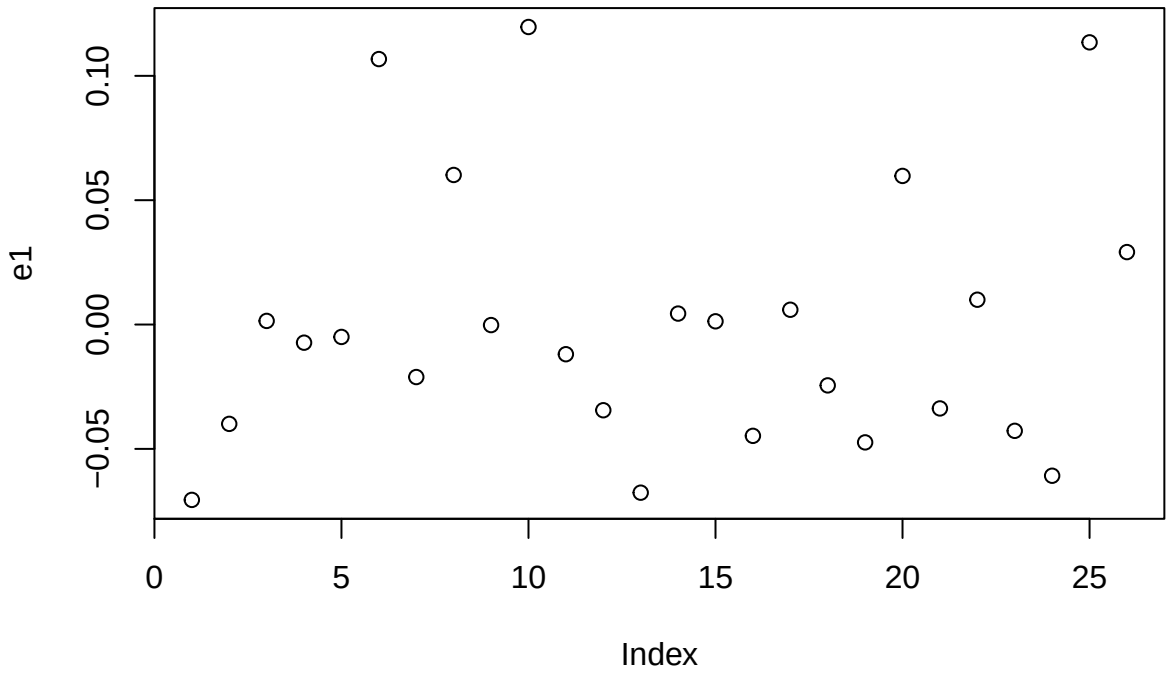
```
##             y         X1         X3
## y   1.0000000  0.7620911 -0.8117565
## X1  0.7620911  1.0000000 -0.4925686
## X3 -0.8117565 -0.4925686  1.0000000
```

- Observation number is not any necessary part for our model so we remove it.
- From this table we can see that X_3 & X_2 have the highest correlation among themselves
- in second table there is a low correlation among the variables.

3.1.10 Extracting residuals and plotting them.

3.1.10.1 residuals of Model 2 we have extracted residuals from the second model and we are plotting them

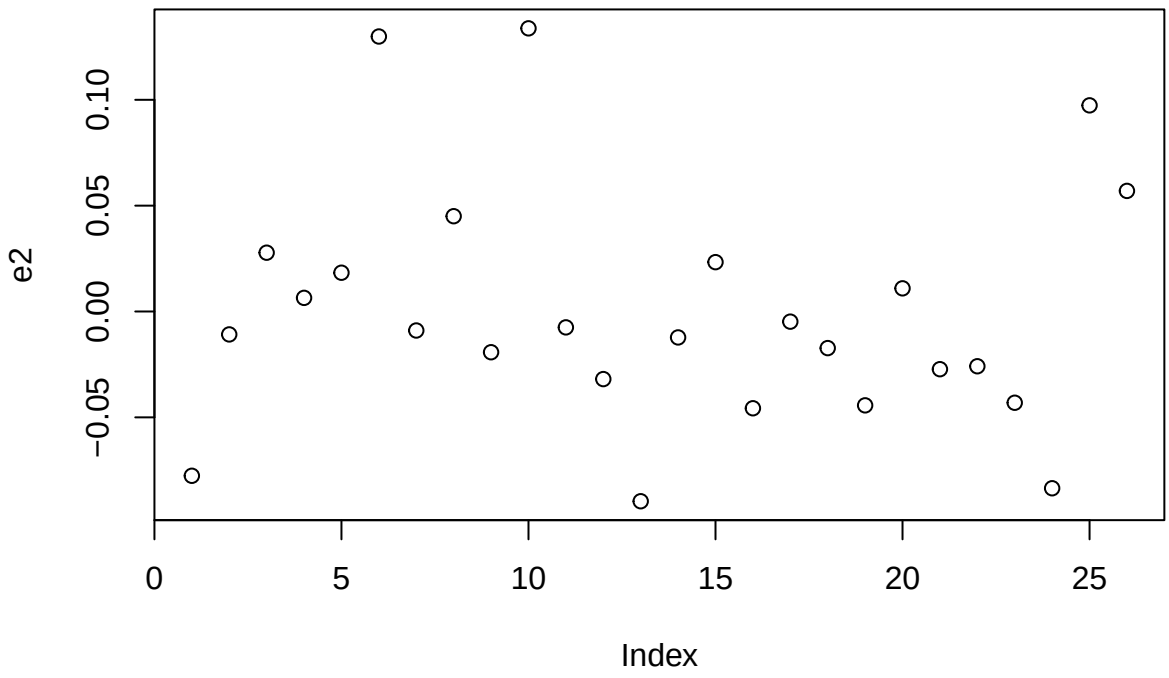
```
e1=sol3$residuals
plot(e1)
```



there is no any pattern shown in this scatter plot of residuals of model 2

3.1.10.2 residual of the best fitted model we have extracted the residuals of model 4 and we are plotting them

```
e2=sol4$residuals
plot(e2)
```



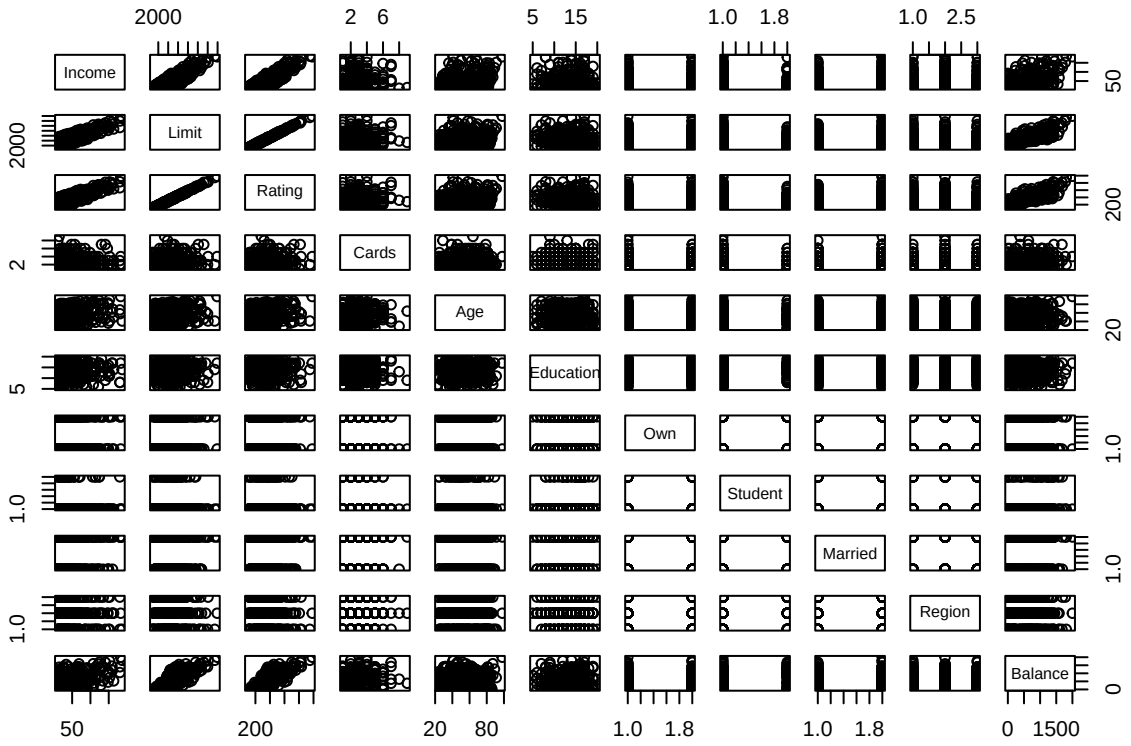
3.2 Assingment 2 & 3

```
credit<-read.csv("credit.csv")
head(credit)
```

```
##    Income Limit Rating Cards Age Education Own Student
## 1  14.891  3606    283     2  34        11  No      No
## 2 106.025  6645    483     3  82        15 Yes     Yes
## 3 104.593  7075    514     4  71        11  No      No
## 4 148.924  9504    681     3  36        11 Yes      No
## 5  55.882  4897    357     2  68        16  No      No
## 6  80.180  8047    569     4  77        10  No      No
##   Married Region Balance
## 1     Yes  South     333
## 2     Yes   West     903
## 3      No   West     580
## 4      No   West     964
## 5     Yes  South     331
## 6      No  South    1151
```

```
own<-factor(credit$Own,labels=c(0,1)) #transforming the (yes,no) values in (0 & 1) i.e integer
student<-factor(credit$Student,labels=c(0,1)) #transforming the (yes,no) values in (0 & 1) i.e integer
married<-factor(credit$Married,labels=c(0,1)) #transforming the (yes,no) values in (0 & 1) i.e integer
```

```
plot(credit)
```



3.2.1 correlation matrix for `credit` data

```
cor(credit[,-7:-10])
```

```
##                Income       Limit      Rating       Cards
## Income     1.00000000  0.79208834  0.79137763 -0.01827261
## Limit      0.79208834  1.00000000  0.99687974  0.01023133
## Rating     0.79137763  0.99687974  1.00000000  0.05323903
## Cards     -0.01827261  0.01023133  0.05323903  1.00000000
## Age        0.17533840  0.10088792  0.10316500  0.04294829
## Education -0.02769198 -0.02354853 -0.03013563 -0.05108422
## Balance    0.46365646  0.86169727  0.86362516  0.08645635
##                   Age    Education      Balance
## Income     0.175338403 -0.027691982  0.463656457
## Limit      0.100887922 -0.023548534  0.861697267
## Rating     0.103164996 -0.030135627  0.863625161
## Cards      0.042948288 -0.051084217  0.086456347
## Age        1.000000000  0.003619285  0.001835119
## Education 0.003619285  1.000000000 -0.008061576
## Balance   0.001835119 -0.008061576  1.000000000
```

- There is high correlation among the variables i.e income & limit,income & rating ,limit & rating.
- As balance is a response variable so there must be a correlation of the variables with them

3.2.2 Setting hypothesis

Null Hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis:
$H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

3.2.2.1 fit the model. $Balance = \beta_0 + \beta_1 \times income + \beta_2 \times limit + \beta_3 \times rating + \beta_4 \times cards + \beta_5 \times age + \beta_6 \times education + \beta_7 \times balance + \epsilon_0$

```
m1<-lm(Balance~Income+Limit+Rating+Cards+Age+Education+own+student+married, data = credit)
summary(m1)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     Education + own + student + married, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -171.66  -75.32  -11.29   54.42  309.98
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -468.40374   34.35512 -13.634  < 2e-16 ***
## Income        -7.80200    0.23395 -33.349  < 2e-16 ***
## Limit          0.19308    0.03268   5.909 7.52e-09 ***
## Rating         1.10227    0.48923   2.253   0.0248 *
## Cards         17.92327    4.33228   4.137 4.31e-05 ***
## Age           -0.63468    0.29325  -2.164   0.0310 *
## Education     -1.11503    1.59592  -0.699   0.4852
## own1         -10.40665    9.90410  -1.051   0.2940
## student1     426.46919   16.67770  25.571  < 2e-16 ***
## married1      -7.01910   10.27803  -0.683   0.4951
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.72 on 390 degrees of freedom
## Multiple R-squared:  0.9549, Adjusted R-squared:  0.9539
## F-statistic: 918.2 on 9 and 390 DF,  p-value: < 2.2e-16
```

- From the above calculated table we came to know that our model is significant as our p-value is less than 0.05
- own and married are insignificant as their p-value is greater than 0.05 except them all the variables are significant.

3.2.3 Setting hypothesis

Null Hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis: $H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

3.2.3.1 fit the model.

$Balance = \beta_0 + \beta_1 \times income + \beta_2 \times limit + \beta_3 \times rating + \beta_4 \times cards + \beta_5 \times age + \beta_7 \times balance + \epsilon_0$

```
m2<-lm(Balance~Income+Limit+Rating+Cards+Age+student, data = credit)

summary(m2)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
##     student, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -170.00  -77.85  -11.84   56.87  313.52
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -493.73419   24.82476 -19.889  < 2e-16 ***
## Income        -7.79508    0.23342 -33.395  < 2e-16 ***
## Limit          0.19369    0.03238   5.981 4.98e-09 ***
## Rating         1.09119    0.48480   2.251   0.0250 *
## Cards         18.21190    4.31865   4.217 3.08e-05 ***
## Age           -0.62406    0.29182  -2.139   0.0331 *
## student1     425.60994   16.50956  25.780  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.61 on 393 degrees of freedom
## Multiple R-squared:  0.9547, Adjusted R-squared:  0.954
## F-statistic:  1380 on 6 and 393 DF,  p-value: < 2.2e-16
```

- our model is significant which implies that parameters has significant effect on the model
- it has the improved adj R-sq and RSE so it is better than the previous one .
- all the parametrs in this case becomes significant after the removal of the education variable from our model.

3.2.4 Durbin watson test for Autocorrelation of the variables

3.2.4.1 Setting hypothesis

Null Hypothesis:

$H_0 : \rho = 0$

VS

Alternative hypothesis: $H_1 : \rho \neq 0$

```
durbinWatsonTest(m2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02182486      1.954429   0.634
##  Alternative hypothesis: rho != 0
```

```
\begin{center}         *OR* \end{center}
```

```
e = m2$residuals
n = length(e) #400
h = numeric(n)
for(i in 2: n)
{
  h[i] = e[i] -e[i-1]
}

#h
d = sum(h^2)/sum(e^2)
d
```

```
## [1] 1.954429
```

as our p-value is greater than 0.05 so we reject the null hypothesis which signifies that there is no autocorrelation among the error terms of the model.Residuals are not autocorrelated.

3.2.5 Correlation of the variables

```
cor(credit[,c(1:6,11)])
```

```
##                Income       Limit      Rating       Cards
## Income     1.00000000  0.79208834  0.79137763 -0.01827261
## Limit      0.79208834  1.00000000  0.99687974  0.01023133
## Rating     0.79137763  0.99687974  1.00000000  0.05323903
## Cards     -0.01827261  0.01023133  0.05323903  1.00000000
## Age        0.17533840  0.10088792  0.10316500  0.04294829
## Education -0.02769198 -0.02354853 -0.03013563 -0.05108422
## Balance    0.46365646  0.86169727  0.86362516  0.08645635
##                   Age    Education      Balance
## Income     0.175338403 -0.027691982  0.463656457
## Limit      0.100887922 -0.023548534  0.861697267
## Rating     0.103164996 -0.030135627  0.863625161
## Cards      0.042948288 -0.051084217  0.086456347
## Age        1.000000000  0.003619285  0.001835119
## Education  0.003619285  1.000000000 -0.008061576
## Balance    0.001835119 -0.008061576  1.000000000
```

- There is high correlation among the variables i.e income & limit,income & rating ,limit & rating.
- As balance is a response variable so there must be a correlation of the variables with them

3.2.6 VIF to check multicolliniarity.

```
vif(m2)
```

```
##     Income      Limit     Rating      Cards        Age
##   2.776906 229.238479 230.869514   1.439007   1.039696
##    student
##   1.009064
```

multicolliniarity is present in this model as rating and limit has very high value of VIF.

3.2.7 After removing Rating

as rating has highest VIF value so it cause huge impact on our model in sense of multicolliniarity

3.2.7.1 Setting hypothesis Null Hypothesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \cdots \quad = \beta_n = 0$

VS

Alternative hypothesis:
$H_1 : At\ least\ one\ of\ the\ \beta_i\ 's\ are\ not\ zero$

3.2.7.1.1 fit the model. $Balance = \beta_0 + \beta_1 \times income + \beta_2 \times limit + \beta_4 \times cards + \beta_5 \times age + \beta_7 \times balance + \epsilon_0$

```
m21<-lm(Balance~Income+Limit+Cards+Age+student, data = credit)
summary(m21)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit + Cards + Age + student,
##     data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -187.05  -79.57  -12.59   56.06  322.56
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.673e+02  2.199e+01 -21.250  < 2e-16 ***
## Income      -7.760e+00  2.341e-01 -33.149  < 2e-16 ***
## Limit        2.661e-01  3.535e-03  75.296  < 2e-16 ***
## Cards        2.355e+01  3.628e+00   6.492 2.55e-10 ***
## Age         -6.220e-01  2.933e-01  -2.120   0.0346 *
## student1     4.284e+02  1.655e+01  25.886  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 99.12 on 394 degrees of freedom
## Multiple R-squared:  0.9541, Adjusted R-squared:  0.9535
## F-statistic:  1638 on 5 and 394 DF,  p-value: < 2.2e-16
```

- In this model all the variables as well as intercept is significant but its Adj R-sq get reduced but it is nearly same to that in model after removing education.

```
vif(m21)
```

3.2.7.2 VIF for

```
##   Income    Limit    Cards      Age  student
## 2.764201 2.703384 1.004962 1.039685 1.003462
```

- VIF is not so high in this case ,hence multicolliniarity is not present.

```
cor(credit[,c(1,2,4,5,6,11)])
```

3.2.7.3 Correlation

```
##                Income       Limit       Cards         Age
## Income     1.00000000  0.79208834 -0.01827261 0.175338403
## Limit      0.79208834  1.00000000  0.01023133 0.100887922
## Cards     -0.01827261  0.01023133  1.00000000 0.042948288
## Age        0.17533840  0.10088792  0.04294829 1.000000000
## Education -0.02769198 -0.02354853 -0.05108422 0.003619285
## Balance    0.46365646  0.86169727  0.08645635 0.001835119
##              Education      Balance
## Income    -0.027691982  0.463656457
## Limit     -0.023548534  0.861697267
## Cards     -0.051084217  0.086456347
## Age        0.003619285  0.001835119
## Education  1.000000000 -0.008061576
## Balance   -0.008061576  1.000000000
```

- Correlation is high only in case of Income and limit but when we remove any of the variable our Adj R-sq value get reduced to large amount.so it is better.

3.2.7.4 Durbin watson test for Autocorrelation of the variables

3.2.7.4.1 Setting hypothesis Null Hypothesis:

$H_0 : \rho = 0$

VS

Alternative hypothesis: $H_1 : \rho \neq 0$

```
durbinWatsonTest(m21)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02182474      1.954855   0.648
##  Alternative hypothesis: rho != 0
```

- it nearly same as that of the previous one . Residuals are not autocorellted.

4 LAB-4

4.1 ASSINGMENT 1

4.1.1 Reading the data from excel

```
household<-read.csv("household.csv")
head(household)
```

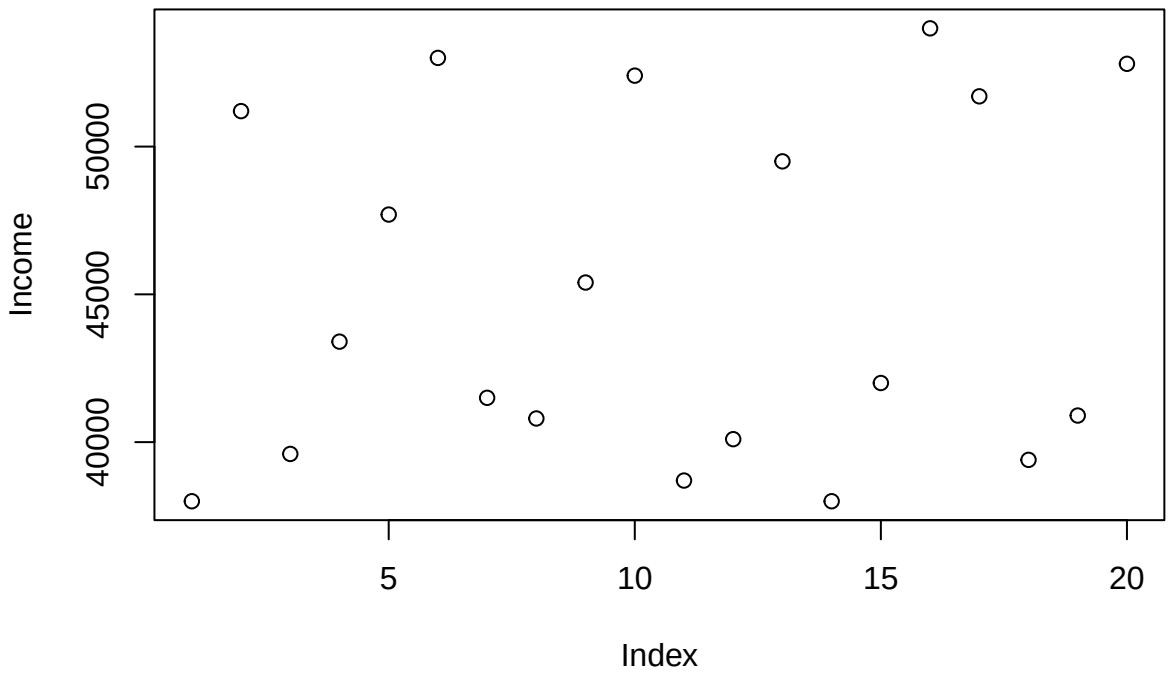
```
##   Household Income Home..Ownership.Status
## 1         1  38000                      0
## 2         2  51200                      1
## 3         3  39600                      0
## 4         4  43400                      1
## 5         5  47700                      0
## 6         6  53000                      0
```

```
dim(household) #dimension of the data household
```

```
## [1] 20  3
```

4.1.2 Plotting the graph

```
attach(household)
plot(Income)
```



Fitting the logistic model

```
logist = glm( Home..Ownership.Status~Income, data= household, family = binomial)
summary(logist)
```

```
##
## Call:
## glm(formula = Home..Ownership.Status ~ Income, family = binomial,
##     data = household)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0232  -0.8766   0.5072   0.7980   1.6046
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.7395139  4.4394326  -1.969   0.0490 *
## Income       0.0002009  0.0001006   1.998   0.0458 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.526  on 19  degrees of freedom
## Residual deviance: 22.435  on 18  degrees of freedom
```

```
## AIC: 26.435
##
## Number of Fisher Scoring iterations: 4
```

4.1.3 Extracting the coefficient from model

```
logist$coef
```

```
##   (Intercept)        Income
## -8.7395139021  0.0002009056
```

4.1.4 predicting probablity

predicting the probablity based on the logistic regression model .

```
logit.prob = predict(logist, type = "response")
logit.prob[1:20]
```

```
##         1         2         3         4         5
## 0.2487856 0.8244590 0.3135336 0.4949479 0.6992408
##         6         7         8         9        10
## 0.8708488 0.4008487 0.3675914 0.5942594 0.8566747
##        11        12        13        14        15
## 0.2759850 0.3355480 0.7694690 0.2487856 0.4251964
##        16        17        18        19        20
## 0.8918125 0.8385268 0.3049509 0.3722741 0.8662619
```

providing *no* to all the predicted values

```
logit.pred = rep("No", 20)
```

assigning *yes* to the probablities that are greater than 0.5

```
logit.pred[logit.prob>0.5]= "Yes" #predicting probablity
logit.pred
```

```
##  [1] "No"  "Yes" "No"  "No"  "Yes" "Yes" "No"  "No"  "Yes"
## [10] "Yes" "No"  "No"  "Yes" "No"  "No"  "Yes" "Yes" "No"
## [19] "No"  "Yes"
```

4.1.5 Confusion matrix

```
v=factor(logit.pred,labels = c(0,1))
table(logit.pred,Home..Ownership.Status)
```

```
##           Home..Ownership.Status
## logit.pred 0 1
##        No  7 4
##        Yes 2 7
```

mean based on the confusion matrix table

```
a=v==Home..Ownership.Status
a
```

```
##  [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE
## [10]  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## [19]  TRUE  TRUE
```

```
mean(v==household$Home..Ownership.Status)
```

```
## [1] 0.7
```

```
#mean
(7+7)/20 #logistic regression correctly predicted the movement of house ownership  70% of the time
```

```
## [1] 0.7
```

4.1.6 Use of simple linear regression model as the structure for linear predictor.

$\eta = log(\frac{\pi}{1-\pi})$ where η is a log of Odd's ratio(log-odd's)and is a liner predictor.
Transformation of $\eta = log(\frac{\pi}{1-\pi})$ is often called *logit transformation* of probablity.

```
beta0=logist$coefficients[1]
beta0
```

```
## (Intercept)
##   -8.739514
```

```
beta1=logist$coefficients[2]
beta1
```

```
##       Income
## 0.0002009056
```

```
estvalue=beta0+beta1*Income
```

here `estvalue` is the estimated value obtained from the simple linear regression model and is the linear predictor of the Home ownership status on the basis of income.

4.1.7 Deviance analysis

```
residual=logist$deviance
residual
```

```
## [1] 22.43492
```

```
resid_df=logist$df.residual
resid_df
```

```
## [1] 18
```

```
dev=22.435/18 # residual/df
dev
```

```
## [1] 1.246389
```

deviance comes out to be *1.2463* as it is close to 1 so it is adequate .

4.1.8 Interpreting the parameter

- β_0 is *-8.73951* and β_1 is *0.0002009* are the coefficients
- β_0 is the intercept term and is negative which led to reduce the home ownership status to some extent.
- β_1 is positive so House ownership status increases with the increase in Income

4.1.9 Adding Quadratic term to the model

```
newmodel=glm(Home..Ownership.Status~Income+I(Income^2),family = binomial)
summary(newmodel)
```

```
##
## Call:
## glm(formula = Home..Ownership.Status ~ Income + I(Income^2),
##     family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9114  -0.8012   0.5864   0.7169   1.8804
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.975e+01  6.105e+01  -1.142    0.253
## Income       2.899e-03  2.682e-03   1.081    0.280
## I(Income^2) -2.940e-08  2.904e-08  -1.012    0.311
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.526  on 19  degrees of freedom
## Residual deviance: 21.326  on 17  degrees of freedom
## AIC: 27.326
##
## Number of Fisher Scoring iterations: 4
```

with the addition of the quadratic term in the model,our model becomes insignificant as all our p values becomes greater than **0.05** .so it is not required to add a quadratic term to the model.

4.2 ASSINGMENGT 2

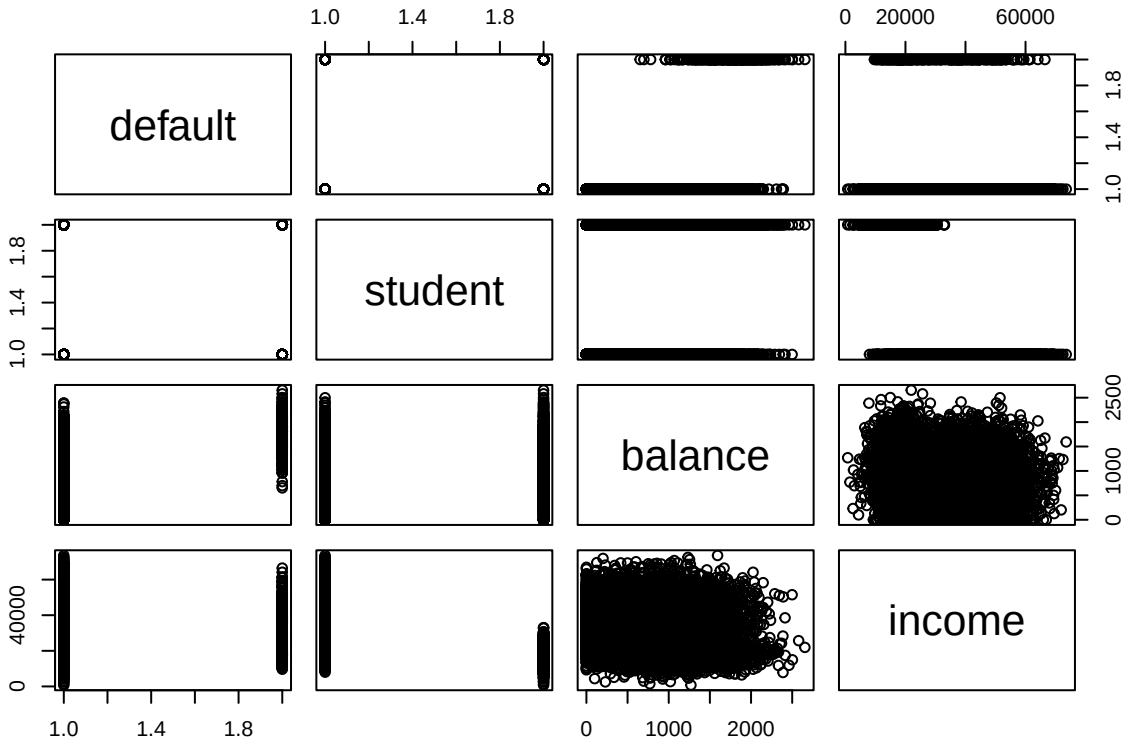
4.2.1 read the deafault data

```
library(ISLR2)
data("Default")
head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

4.2.2 Plotting the data

```
pairs(Default)
```



there is no linear relationship between any variables

4.2.3 Fitting the model

```
mod= glm(default~student+balance+income,data = Default,family = binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = default ~ student + balance + income, family = binomial,
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

all the variables are significant excpt the `income` as its p-value is greater than 0.05 so we will remove it from our model.

```
mod1= glm(default~student+balance,data = Default,family = binomial)
summary(mod1)
```

```
##
## Call:
## glm(formula = default ~ student + balance, family = binomial,
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4578  -0.1422  -0.0559  -0.0203   3.7435
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
## studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
## balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
```

```
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.7  on 9997  degrees of freedom
## AIC: 1577.7
##
## Number of Fisher Scoring iterations: 8
```

its `AIC` value is less than the previous model so it is better than that model

predicting the values for the `mod1`

```
pred.mod=predict(mod1,type = "response")
pred.mod[1:20]
```

```
##            1            2            3            4
## 1.409096e-03 1.140318e-03 1.005719e-02 4.469571e-04
##            5            6            7            8
## 1.943498e-03 2.050378e-03 2.441704e-03 1.086013e-03
##            9           10           11           12
## 1.650864e-02 2.145576e-05 1.049739e-05 1.142414e-02
##           13           14           15           16
## 8.360645e-05 6.970577e-04 1.257776e-02 1.108675e-04
##           17           18           19           20
## 2.145576e-05 2.165889e-04 3.486339e-04 1.136424e-02
```

assingning yes to the values having probablity greater than **0.5** .

```
pred.class=rep("No",10000) # assigning __no__ to all the probablities
pred.class[pred.mod>0.5]="Yes"  # assigning __yes__ to the probablities greater than 0.5
pred.class[1:20]
```

```
##  [1] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
## [11] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
```

4.2.4 confusion matrix

```
default=Default$default
table(pred.class,default)
```

```
##           default
## pred.class   No  Yes
##        No  9628  228
##        Yes   39  105
```

```
mean(pred.class==default)
```

```
## [1] 0.9733
```

```
(9628+105)/10000 #logistic regression correctly predicted the movement of default data 97.33% of the ti
```

```
## [1] 0.9733
```

4.2.4.1 Fitting the model using Validation Set Approach (or Train and Test data)

dividing the data into test and and train data

```
balance <- Default$balance
dat <- (income<40000)
trainData <- Default[dat,]
dim(trainData)
```

```
## [1] 6503    4
```

```
testData <- Default[!dat,]
dim(testData)
```

```
## [1] 3497    4
```

```
testDefault <- default[!dat]
```

```
mod.fit = glm(default ~ student + balance,data = Default,family ='binomial',subset = dat) #fitting our
summary(mod.fit)
```

4.2.4.1.1 fitting the model based on train and test data

```
##
## Call:
## glm(formula = default ~ student + balance, family = "binomial",
##     data = Default, subset = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1517  -0.1491  -0.0581  -0.0216   3.6633
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.069e+01  4.504e-01 -23.727  < 2e-16 ***
## studentYes  -7.039e-01  1.718e-01  -4.096  4.2e-05 ***
## balance      5.696e-03  2.809e-04  20.275  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1975.8  on 6502  degrees of freedom
## Residual deviance: 1072.4  on 6500  degrees of freedom
## AIC: 1078.4
##
## Number of Fisher Scoring iterations: 8
```

```
mod.prob = predict(mod.fit,testData, type = 'response') #predicting probablities

mod.pred = rep("No", 3497)
mod.pred[mod.prob>0.5]= "Yes"
head(mod.pred)
```

```
## [1] "No" "No" "No" "No" "No" "No"
```

```
table(mod.pred, testDefault)
```

```
##         testDefault
## mod.pred   No  Yes
##      No  3381   69
##      Yes   11   36
```

```
mean(mod.pred == testDefault)
```

```
## [1] 0.9771232
```

```
mean(mod.pred != testDefault)
```

```
## [1] 0.02287675
```

```
(3381+36)/3497 #logistic regression correctly predicted the movement of default data 97.71% of the time
```

```
## [1] 0.9771232
```

previously the model has the accuracy of *0.9733* and in validation set approach accuracy has been increased *0.9771*.

5 LAB-5

5.1 Assingment-1

5.1.1 Reading the data from excel

```
household<-read.csv("household.csv")
head(household)
```

```
##   Household Income Home..Ownership.Status
## 1         1  38000                      0
## 2         2  51200                      1
## 3         3  39600                      0
## 4         4  43400                      1
## 5         5  47700                      0
## 6         6  53000                      0
```

In R we will fit the model using `lda()` function, which is in the MASS library. Divide the data in the train and test set and calculate the `mse` .

```
attach(household)
```

```
## The following objects are masked from household (pos = 3):
##
##     Home..Ownership.Status, Household, Income
```

```
train=(Income<50500)
trainData=household[train,]
testData=household[!train,]
dim(trainData) #dimension of trainData
```

```
## [1] 14  3
```

```
dim(testData) #dimension of testData
```

```
## [1] 6 3
```

```
testOwnershipStatus=Home..Ownership.Status[!train]
```

5.1.2 Fitting the LDA model

we are fitting the LDA model using `lda()`

```


require(MASS)
```

5.1.2.1 random guessing

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     chem
```

```
## The following object is masked from 'package:wooldridge':
##
##     cement

## The following object is masked from 'package:ISLR2':
##
##     Boston
```

```
ownership_status.fit_rguess<-lda(Home..Ownership.Status~Income,data = household)
ownership_status.fit_rguess
```

```
## Call:
## lda(Home..Ownership.Status ~ Income, data = household)
##
## Prior probabilities of groups:
##    0    1
## 0.45 0.55
##
## Group means:
##     Income
## 0 41944.44
## 1 47509.09
##
## Coefficients of linear discriminants:
##                 LD1
## Income 0.0001899588
```

```
ownership_status.pred_rguess<-predict(ownership_status.fit_rguess,household)
ownership_status.class_rguess=ownership_status.pred_rguess$class
table(ownership_status.class_rguess,household$Home..Ownership.Status) #confusion matrix based on random
```

```
##
## ownership_status.class_rguess 0 1
##                             0 7 4
##                             1 2 7
```

```
mean_rguess=mean(ownership_status.class_rguess==household$Home..Ownership.Status)  #mean based on rando
mean_rguess
```

```
## [1] 0.7
```

```
require(MASS)
ownership_status.fit<-lda(Home..Ownership.Status~Income,data = household,subset = train)
ownership_status.fit
```

5.1.2.1.1 validation set approach

```
## Call:
## lda(Home..Ownership.Status ~ Income, data = household, subset = train)
##
## Prior probabilities of groups:
##         0         1
## 0.5714286 0.4285714
##
## Group means:
##     Income
## 0 40562.50
## 1 43416.67
##
## Coefficients of linear discriminants:
##                 LD1
## Income 0.0002974743
```

• The LDA output indicates that $\hat{\pi}_1$ =0.5714286 and $\hat{\pi}_2$ =0.4285714 ,there are 75%% of the training observation corresponding to the persons who doesn't have the Home ownership.

• It also provides the group means,these are the average of each predictor within each class,and are used by LDA as estimates of μ_k. This suggest that income have greater influence on the House ownership **(43416.67)** than the non house ownership **(40562.50)**.

• The coefficient of linear discriminant output provides the linear combination of Income and Household that are used to form the LDA descision rule.
0.0002974743 $\times$ **Income**

this line represent the boundary between the house ownership and non house ownership.Home ownership will be predicted depending upon the which side of the line they are.

5.1.3 use the model to make predictions

as we have fit the model using our training data, now we use it to make predictions on our test data .

```
ownership_status.pred<-predict(ownership_status.fit,testData)
names(ownership_status.pred)
```

```
## [1] "class"     "posterior" "x"
```

```
ownership_status.class<-ownership_status.pred$class
```

it contain a list of three variables *Class*:the predicted class *posterior*:the posterior probability that an observation belong to each class x: the linear discriminant

5.1.4 Confusion matrix

```
table(ownership_status.class,testOwnershipStatus) # confusion matrix
```

```
##                       testOwnershipStatus
## ownership_status.class 0 1
##                      0 0 0
##                      1 1 5
```

```
mean_vsa =mean(ownership_status.class==testOwnershipStatus)
mean_vsa
```

```
## [1] 0.8333333
```

it turns out that model predicted the testOwnershipStatus for 83.33% of the observation in our data correctly.

Applying a 9% threshold to the posterior probabilities allow us to recreate the predictions contained in `ownership_status.pred$class`.

```
sum(ownership_status.pred$posterior[,1]>=.09)
```

```
## [1] 2
```

```
sum(ownership_status.pred$posterior[,1]<.09)
```

```
## [1] 4
```

```
ownership_status.pred$posterior[1:6,1]
```

```
##          2          6         10         16         17
## 0.11521204 0.07633654 0.08773134 0.06032633 0.10295096
##         20
## 0.07997532
```

```
ownership_status.class[1:6]
```

```
## [1] 1 1 1 1 1 1
## Levels: 0 1
```

there are 2 persons those are the non house owners and have their posterior probability greater than 9%.

if we wanted to use a posterior probability threshold other than 50% in order to make prediction.

```
sum(ownership_status.pred$posterior[,1]>.12)
```

```
## [1] 0
```

No person meet that threshold which implies that there is no person that doesn't have the house ownership and has threshold greater than 12%.

5.1.5 Conclusion

Mean_rguess(*70.00%*) of the Household ownership in case of validation set approach is grater than the Mean_vsa (*81.25%*) in case of validation set approach

5.2 Assingment 2

Use default data and divide the data in train and test

```
require(ISLR2)
data("Default")
attach(Default)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     balance, default, income, student
```

```
head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

```
train=(income<40000)
trainData=Default[train,]
dim(trainData)
```

```
## [1] 6503    4
```

```
testData=Default[!train,]
dim(testData)
```

```
## [1] 3497    4
```

```
testDefault=default[!train]
```

5.2.1 fit the model using `lda()`

fit the data using `lda()` function based on the train data.

```
require(MASS)

default_lda.fit_rguess<-lda(default~student+balance+income,data = Default)
default_lda.fit_rguess
```

5.2.1.1 random guessing

```
## Call:
## lda(default ~ student + balance + income, data = Default)
##
## Prior probabilities of groups:
##     No    Yes
```

```
## 0.9667 0.0333
##
## Group means:
##     studentYes   balance   income
## No   0.2914037  803.9438 33566.17
## Yes  0.3813814 1747.8217 32089.15
##
## Coefficients of linear discriminants:
##                       LD1
## studentYes -1.746631e-01
## balance     2.243541e-03
## income      3.367310e-06
```

```
default_lda.pred_rguess<-predict(default_lda.fit_rguess,Default)
default_lda.class_rguess=default_lda.pred_rguess$class
table(default_lda.class_rguess,default) #confusion matrix based on random guessing
```

```
##                         default
## default_lda.class_rguess   No  Yes
##                       No  9645  254
##                       Yes   22   79
```

```
mean_rguess=mean(default_lda.class_rguess==default)  #mean based on random guessing
mean_rguess
```

```
## [1] 0.9724
```

```
default_lda.fit<-lda(default~student+balance+income,data = Default,subset = train)
default_lda.fit
```

5.2.1.2 validation set approach

```
## Call:
## lda(default ~ student + balance + income, data = Default, subset = train)
##
## Prior probabilities of groups:
##         No        Yes
## 0.96493926 0.03506074
##
## Group means:
##     studentYes   balance   income
## No   0.4489243  834.8563 25730.68
## Yes  0.5570175 1771.1333 24409.13
##
## Coefficients of linear discriminants:
##                       LD1
## studentYes -2.019107e-01
## balance     2.235850e-03
## income      3.352984e-06
```

• The LDA output indicates that $\hat{\pi}_1$ =0.96493926 and $\hat{\pi}_2$ =0.03506074 ,there are 96.49% of the training observation corresponding to the persons who doesn't have been defaulted by debt.

• It also provides the group means,these are the average of each predictor within each class,and are used by LDA as estimates of μ_k. This suggest that student have greater influence on the the customer defaulted on their debt **(0.5570175)** than the non student **(0.4489243)** similarly balance has greater influence on the defaulter **(1771.1333)** than the non defaulter **(834.8563)** but in case of the income it is opposite as greater income **(25730.68)** are not the defaulter but slightly low income **(24409.13)** are the defaulter

• The coefficient of linear discriminant output provides the linear combination of studentYes ,balance and income that are used to form the LDA descision rule.
** -2.019107e-01 × studentYes + 2.235850e-03 × balance + 3.352984e-06 × income ** this plane represent the boundary between the deafault and non dafault.default will be predicted depending upon the which side of the line they are.

5.2.2 predicting the values

predict the values based on the test data

```
default_lda.pred<-predict(default_lda.fit,testData)
names(default_lda.pred)
```

```
## [1] "class"     "posterior" "x"
```

it contain a list of three variables *Class*:the predicted class *posterior*:the posterior probability that an observation belong to each class x: the linear discriminant

5.2.3 confusion matrix

```
default_lda.class=default_lda.pred$class
table(default_lda.class,testDefault)
```

```
##                   testDefault
## default_lda.class   No  Yes
##               No  3385   79
##               Yes    7   26
```

```
mean_vsa=mean(default_lda.class==testDefault)
mean_vsa
```

```
## [1] 0.9754075
```

Applying a 50% threshold to the posterior probabilities allow us to recreate the predictions contained in `default_lda.pred$class`.

```
sum(default_lda.pred$posterior[,1]>=0.5)
```

```
## [1] 3464
```

```
sum(default_lda.pred$posterior[,1]<0.5)
```

```
## [1] 33
```

there are 3464 persons that are defaulted by debt having the posterior probability greater than the threshold of 50% and there are 33 persons that are defaulted by debt having posterior probability less than than the threshold of 50%.

```
default_lda.pred$posterior[1:20,1]
```

```
##         1        14        16        17        19
## 0.9963949 0.9979439 0.9995317 0.9998706 0.9986794
##        21        23        27        31        34
## 0.9996267 0.9831247 0.9976281 0.9986909 0.9914518
##        37        41        42        43        46
## 0.9968936 0.9931763 0.9937851 0.9754198 0.9991211
##        47        49        51        55        58
## 0.9619248 0.9943808 0.9983420 0.9998709 0.8654330
```

```
default_lda.class[1:20]
```

```
##  [1] No No No No No No No No No No No No No No No No No No No No
## [19] No No
## Levels: No Yes
```

```
sum(default_lda.pred$posterior[,1]>.99)
```

```
## [1] 2264
```

there are 2264 persons that are defaulted that have the posterior probability greater than the 99% threshold .

5.2.4 comparision between both the result

In case of random guessing data the mean was **0.9724** but in case of validation set approach mean is **0.9754075** validation set approach data has the higher mean as compared to the data based on the random guessing.

6 LAB-6

6.1 QDA-Quadratic Discriminant Analysis

6.1.1 fiting the model using `qda()`

```
require(MASS)
data("Default")

default_qda.fit_rguess<-qda(default~student+balance+income,data = Default)
default_qda.fit_rguess
```

6.1.1.1 random guessing

```
## Call:
## qda(default ~ student + balance + income, data = Default)
##
## Prior probabilities of groups:
##     No    Yes
## 0.9667 0.0333
##
## Group means:
##     studentYes   balance   income
## No   0.2914037  803.9438 33566.17
## Yes  0.3813814 1747.8217 32089.15
```

```
default_qda.pred_rguess<-predict(default_qda.fit_rguess,Default)
default_qda.class_rguess=default_qda.pred_rguess$class
table(default_qda.class_rguess,default) #confusion matrix based on random guessing
```

```
##                         default
## default_qda.class_rguess   No  Yes
##                      No  9636  239
##                      Yes   31   94
```

```
mean_rguess=mean(default_qda.class_rguess==default)  #mean based on random guessing
mean_rguess
```

```
## [1] 0.973
```

```
default_qda.fit<-qda(default~.,data=Default,subset = train)
default_qda.fit
```

6.1.1.2 validation set approach

```
## Call:
## qda(default ~ ., data = Default, subset = train)
##
## Prior probabilities of groups:
##         No        Yes
## 0.96493926 0.03506074
##
## Group means:
```

```
##     studentYes   balance   income
## No   0.4489243  834.8563 25730.68
## Yes  0.5570175 1771.1333 24409.13
```

the output returns the posterior probabilities and group means as LDA, except the linear discriminant because the QDA classifier involves a quadratic, rather than a linear, function of the predictors.

6.1.2 predicting the data

```
default_qda.pred<-predict(default_qda.fit,testData)
default_qda.class<-default_qda.pred$class
table(default_qda.class,testDefault) # confusion matrix
```

```
##                  testDefault
## default_qda.class   No  Yes
##               No  3388   89
##               Yes    4   16
```

```
mean_vsa=mean(default_qda.class==testDefault) #mean
```

the QDA mean_vsa **0.9734** in case of the validation set approach but in case of the random guessing the mean_rguess is **0.973** . it has lower mean than the mean_rguess of LDA **0.9754075**

7 LAB-7

7.1 Naive Bayes

7.1.1 fit the model using `naiveBayes()`

```
library(e1071)
library(naivebayes)
```

```
## naivebayes 0.9.7 loaded
```

```
data("Default")

default_naive.fit_rguess<-naive_bayes(default~student+balance+income,data = Default)
default_naive.fit_rguess
```

```
##
## ====================== Naive Bayes ======================
##
##  Call:
## naive_bayes.formula(formula = default ~ student + balance + income,
##     data = Default)
##
## ---------------------------------------------------------
```

```
##
## Laplace smoothing: 0
##
## -------------------------------------------------------------
##
##  A priori probabilities:
##
##     No    Yes
## 0.9667 0.0333
##
## -------------------------------------------------------------
##
##  Tables:
##
## -------------------------------------------------------------
##  ::: student (Bernoulli)
## -------------------------------------------------------------
##
## student        No       Yes
##     No  0.7085963 0.6186186
##     Yes 0.2914037 0.3813814
##
## -------------------------------------------------------------
##  ::: balance (Gaussian)
## -------------------------------------------------------------
##
## balance        No       Yes
##    mean  803.9438 1747.8217
##    sd    456.4762  341.2668
##
## -------------------------------------------------------------
##  ::: income (Gaussian)
## -------------------------------------------------------------
##
## income       No      Yes
##   mean 33566.17 32089.15
##   sd   13318.25 13804.22
##
## -------------------------------------------------------------
```

```
default_naive.pred_rguess<-predict(default_naive.fit_rguess,Default)

default_naive.class_rguess=default_naive.pred_rguess #as it gives only values in YES and NO

table(default_naive.class_rguess,default) #confusion matrix based on random guessing
```

```
##                          default
## default_naive.class_rguess   No  Yes
##                        No  9615  241
##                        Yes   52   92
```

```
mean_rguess=mean(default_naive.class_rguess==default)  #mean based on random guessing
mean_rguess
```

```
## [1] 0.9707
```

```
default_nb.fit<-naive_bayes(default~.,data= Default,subset = train)
default_nb.fit
```

7.1.1.1 validation set approach

```
##
## ====================== Naive Bayes ======================
##
##  Call:
## naive_bayes.formula(formula = default ~ ., data = Default, subset = train)
##
## ---------------------------------------------------------
##
## Laplace smoothing: 0
##
## ---------------------------------------------------------
##
##  A priori probabilities:
##
##         No        Yes
## 0.96493926 0.03506074
##
## ---------------------------------------------------------
##
##  Tables:
##
## ---------------------------------------------------------
##  ::: student (Bernoulli)
## ---------------------------------------------------------
##
## student        No       Yes
##     No  0.5510757 0.4429825
##     Yes 0.4489243 0.5570175
##
## ---------------------------------------------------------
##  ::: balance (Gaussian)
## ---------------------------------------------------------
##
## balance        No       Yes
##    mean  834.8563 1771.1333
##    sd    460.1149  336.3134
##
## ---------------------------------------------------------
##  ::: income (Gaussian)
## ---------------------------------------------------------
##
## income        No       Yes
##   mean 25730.680 24409.128
##   sd    8878.337  8631.365
```

```
##
## ------------------------------------------------------------
```

the output contains the estimated mean and standard deviation for each variable in the group. Mean of the income is 25730.680 for no default and standard deviation is 8878.337 similarly mean of the balance is 834.8563 for no default and standard deviation is 460.1149

7.1.2 finding mean and standard deviation for balance

```
mean(balance[train][default[train]=="No"])
```

```
## [1] 834.8563
```

```
sd(balance[train][default[train]=="No"])
```

```
## [1] 460.1149
```

7.1.3 predicting the values based on the test data

```
default_nb.pred<-predict(default_nb.fit,testData)
table(default_nb.pred,testDefault) #confusion matrix
```

```
##                testDefault
## default_nb.pred   No  Yes
##             No  3391  102
##             Yes    1    3
```

```
mean(default_nb.pred==testDefault) #mean
```

```
## [1] 0.9705462
```

Naive Bayes performs very well with accurate predictions over 97.05% of the time ,This is slightly worse than the QDA (*97.34%*) and LDA (*97.54*).