Advanced Linear Regression

#Importing libraries  
library(readxl)  
library(corrplot)

## corrplot 0.84 loaded

library(stargazer)

## Warning: package 'stargazer' was built under R version 4.0.3

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

setwd('C:/Users/surya/Downloads')  
cr = read\_excel("CreditRating.xlsx", sheet = 'Data')  
  
#NA values column wise  
sapply(cr, function(x) sum(is.na(x)))

## ID Income Limit Rating Cards Age Education Gender   
## 0 0 0 0 0 0 0 0   
## Student Married Ethnicity Balance   
## 0 0 0 0

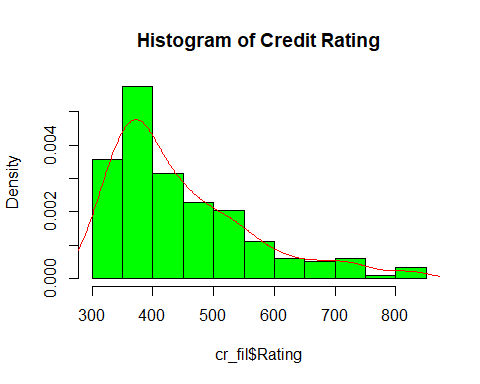
str(cr)

## tibble [400 x 12] (S3: tbl\_df/tbl/data.frame)  
## $ ID : num [1:400] 1 2 3 4 5 6 7 8 9 10 ...  
## $ Income : num [1:400] 14.9 106 104.6 148.9 55.9 ...  
## $ Limit : num [1:400] 3606 6645 7075 9504 4897 ...  
## $ Rating : num [1:400] 283 483 514 681 357 569 259 512 266 491 ...  
## $ Cards : num [1:400] 2 3 4 3 2 4 2 2 5 3 ...  
## $ Age : num [1:400] 34 82 71 36 68 77 37 87 66 41 ...  
## $ Education: num [1:400] 11 15 11 11 16 10 12 9 13 19 ...  
## $ Gender : chr [1:400] "Male" "Female" "Male" "Female" ...  
## $ Student : chr [1:400] "No" "Yes" "No" "No" ...  
## $ Married : chr [1:400] "Yes" "Yes" "No" "No" ...  
## $ Ethnicity: chr [1:400] "Caucasian" "Hispanic" "Hispanic" "Hispanic" ...  
## $ Balance : num [1:400] 333 903 580 964 331 ...

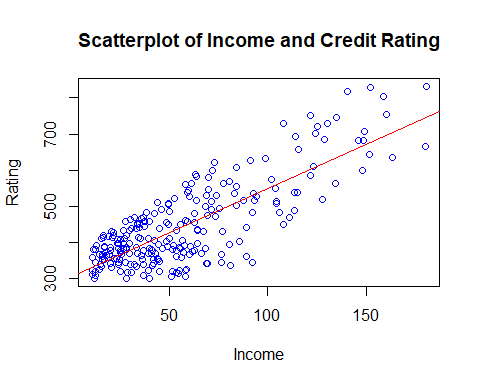
#Basic statistics  
summary(cr)

## ID Income Limit Rating   
## Min. : 1.0 Min. : 10.35 Min. : 855 Min. : 93.0   
## 1st Qu.:100.8 1st Qu.: 21.01 1st Qu.: 3088 1st Qu.:247.2   
## Median :200.5 Median : 33.12 Median : 4622 Median :344.0   
## Mean :200.5 Mean : 45.22 Mean : 4736 Mean :354.9   
## 3rd Qu.:300.2 3rd Qu.: 57.47 3rd Qu.: 5873 3rd Qu.:437.2   
## Max. :400.0 Max. :186.63 Max. :13913 Max. :982.0   
## Cards Age Education Gender   
## Min. :1.000 Min. :23.00 Min. : 5.00 Length:400   
## 1st Qu.:2.000 1st Qu.:41.75 1st Qu.:11.00 Class :character   
## Median :3.000 Median :56.00 Median :14.00 Mode :character   
## Mean :2.958 Mean :55.67 Mean :13.45   
## 3rd Qu.:4.000 3rd Qu.:70.00 3rd Qu.:16.00   
## Max. :9.000 Max. :98.00 Max. :20.00   
## Student Married Ethnicity Balance   
## Length:400 Length:400 Length:400 Min. : 0.00   
## Class :character Class :character Class :character 1st Qu.: 68.75   
## Mode :character Mode :character Mode :character Median : 459.50   
## Mean : 520.01   
## 3rd Qu.: 863.00   
## Max. :1999.00

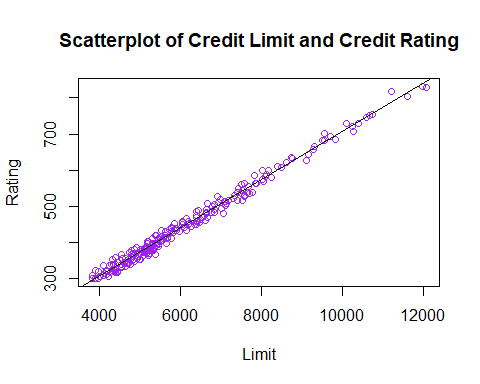
#Credit rating can be between 300 and 850  
cr\_fil = cr[!(cr$Rating < 300 | cr$Rating > 850),]  
  
hist(cr\_fil$Rating, prob=T, main = "Histogram of Credit Rating", col = 'green')  
density\_1 = density(cr\_fil$Rating)   
lines(density\_1, col="red")



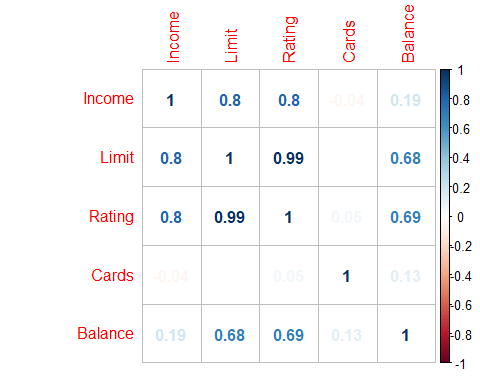
plot(cr\_fil$Income, cr\_fil$Rating, main='Scatterplot of Income and Credit Rating',  
 col = 'blue', xlab = "Income", ylab='Rating')  
abline(lm(Rating ~ Income, data = cr\_fil), col = "red")



plot(cr\_fil$Limit, cr\_fil$Rating, main='Scatterplot of Credit Limit and Credit Rating',  
 col = 'purple', xlab = "Limit", ylab='Rating')  
abline(lm(Rating ~ Limit, data = cr\_fil), col = "black")



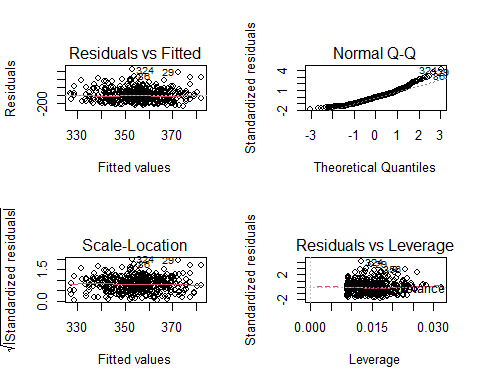
#Checking correlation  
cr1 = subset(cr\_fil, select = c(Income, Limit, Rating, Cards, Balance))  
cr1\_corr = cor(cr1)  
corrplot(cr1\_corr, method = "number")



#Credit rating and limit are directly proportional  
  
#1st linear model  
m1 <- lm(Rating ~ Gender + Education + Married + Ethnicity, data=cr)  
summary(m1)

##   
## Call:  
## lm(formula = Rating ~ Gender + Education + Married + Ethnicity,   
## data = cr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -271.23 -106.04 -8.34 81.19 629.16   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 380.502 38.490 9.886 <2e-16 \*\*\*  
## GenderMale -2.896 15.548 -0.186 0.852   
## Education -1.578 2.493 -0.633 0.527   
## MarriedYes 13.987 16.072 0.870 0.385   
## EthnicityCaucasian -11.923 19.169 -0.622 0.534   
## EthnicityHispanic -21.860 22.071 -0.990 0.323   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 155.3 on 394 degrees of freedom  
## Multiple R-squared: 0.004925, Adjusted R-squared: -0.007703   
## F-statistic: 0.39 on 5 and 394 DF, p-value: 0.8556

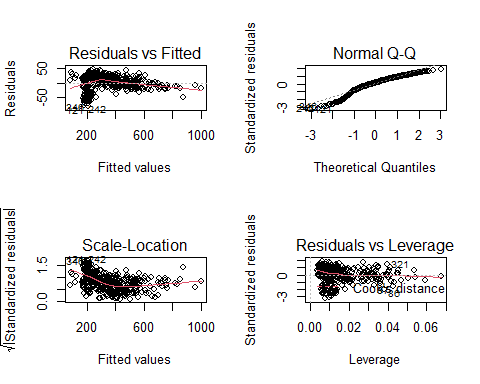
par(mfrow = c(2, 2))  
plot(m1)



par(mfrow=c(1,1))  
  
#Residuals seems to follow an inverted U pattern which implies non-linearity  
#2nd linear model  
m2 <- lm(Rating ~ Income + Cards + Age + Student + Balance, data = cr)  
summary(m2)

##   
## Call:  
## lm(formula = Rating ~ Income + Cards + Age + Student + Balance,   
## data = cr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79.807 -11.074 4.012 18.111 45.823   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 144.095510 5.095584 28.279 <2e-16 \*\*\*  
## Income 2.064114 0.041228 50.065 <2e-16 \*\*\*  
## Cards -0.438944 0.919925 -0.477 0.6335   
## Age 0.124793 0.074023 1.686 0.0926 .   
## StudentYes -98.421649 4.345393 -22.650 <2e-16 \*\*\*  
## Balance 0.234034 0.003235 72.354 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.96 on 394 degrees of freedom  
## Multiple R-squared: 0.9743, Adjusted R-squared: 0.974   
## F-statistic: 2986 on 5 and 394 DF, p-value: < 2.2e-16

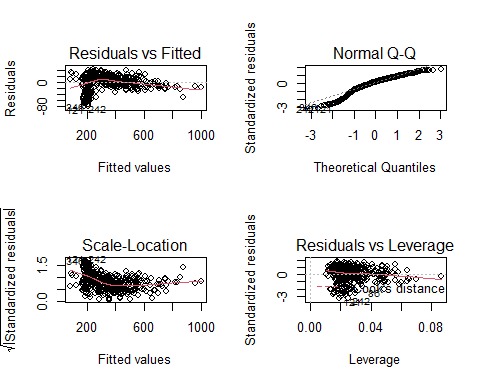
par(mfrow = c(2, 2))  
plot(m2)



par(mfrow=c(1,1))  
  
#3rd linear model  
m3 <- lm(Rating ~ Income + Cards + Age + Student + Balance + Gender + Education + Married + Ethnicity, data = cr)  
summary(m3)

##   
## Call:  
## lm(formula = Rating ~ Income + Cards + Age + Student + Balance +   
## Gender + Education + Married + Ethnicity, data = cr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79.76 -11.54 5.07 17.00 43.08   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 144.879926 8.118984 17.845 <2e-16 \*\*\*  
## Income 2.061675 0.041298 49.922 <2e-16 \*\*\*  
## Cards -0.382905 0.920745 -0.416 0.6777   
## Age 0.124232 0.074381 1.670 0.0957 .   
## StudentYes -97.995439 4.380915 -22.369 <2e-16 \*\*\*  
## Balance 0.233929 0.003234 72.332 <2e-16 \*\*\*  
## GenderMale -2.947742 2.502496 -1.178 0.2395   
## Education 0.079960 0.402317 0.199 0.8426   
## MarriedYes 3.683403 2.601197 1.416 0.1576   
## EthnicityCaucasian -2.506325 3.083337 -0.813 0.4168   
## EthnicityHispanic -5.722754 3.557106 -1.609 0.1085   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.95 on 389 degrees of freedom  
## Multiple R-squared: 0.9747, Adjusted R-squared: 0.974   
## F-statistic: 1496 on 10 and 389 DF, p-value: < 2.2e-16

par(mfrow = c(2, 2))  
plot(m3)



par(mfrow=c(1,1))  
  
#Stargazer  
stargazer(m1, m2, type='text', single.row = TRUE)

##   
## =====================================================================  
## Dependent variable:   
## --------------------------------------  
## Rating   
## (1) (2)   
## ---------------------------------------------------------------------  
## GenderMale -2.896 (15.548)   
## Education -1.578 (2.493)   
## MarriedYes 13.987 (16.072)   
## EthnicityCaucasian -11.923 (19.169)   
## EthnicityHispanic -21.860 (22.071)   
## Income 2.064\*\*\* (0.041)   
## Cards -0.439 (0.920)   
## Age 0.125\* (0.074)   
## StudentYes -98.422\*\*\* (4.345)  
## Balance 0.234\*\*\* (0.003)   
## Constant 380.502\*\*\* (38.490) 144.096\*\*\* (5.096)  
## ---------------------------------------------------------------------  
## Observations 400 400   
## R2 0.005 0.974   
## Adjusted R2 -0.008 0.974   
## Residual Std. Error (df = 394) 155.319 24.965   
## F Statistic (df = 5; 394) 0.390 2,986.402\*\*\*   
## =====================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

stargazer(m1, m2, type='text', ci=TRUE, ci.level=0.95, single.row = TRUE)

##   
## ===========================================================================================  
## Dependent variable:   
## ------------------------------------------------------------  
## Rating   
## (1) (2)   
## -------------------------------------------------------------------------------------------  
## GenderMale -2.896 (-33.369, 27.578)   
## Education -1.578 (-6.464, 3.308)   
## MarriedYes 13.987 (-17.513, 45.488)   
## EthnicityCaucasian -11.923 (-49.493, 25.647)   
## EthnicityHispanic -21.860 (-65.119, 21.398)   
## Income 2.064\*\*\* (1.983, 2.145)   
## Cards -0.439 (-2.242, 1.364)   
## Age 0.125\* (-0.020, 0.270)   
## StudentYes -98.422\*\*\* (-106.938, -89.905)  
## Balance 0.234\*\*\* (0.228, 0.240)   
## Constant 380.502\*\*\* (305.064, 455.940) 144.096\*\*\* (134.108, 154.083)   
## -------------------------------------------------------------------------------------------  
## Observations 400 400   
## R2 0.005 0.974   
## Adjusted R2 -0.008 0.974   
## Residual Std. Error (df = 394) 155.319 24.965   
## F Statistic (df = 5; 394) 0.390 2,986.402\*\*\*   
## ===========================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

m = mean(cr$Cards)  
m

## [1] 2.9575

sd(cr$Cards)

## [1] 1.371275

err = 0.918  
  
lb = err - m  
lb

## [1] -2.0395

ub = err + m  
ub

## [1] 3.8755

NROW(cr$Cards)

## [1] 400