regression project
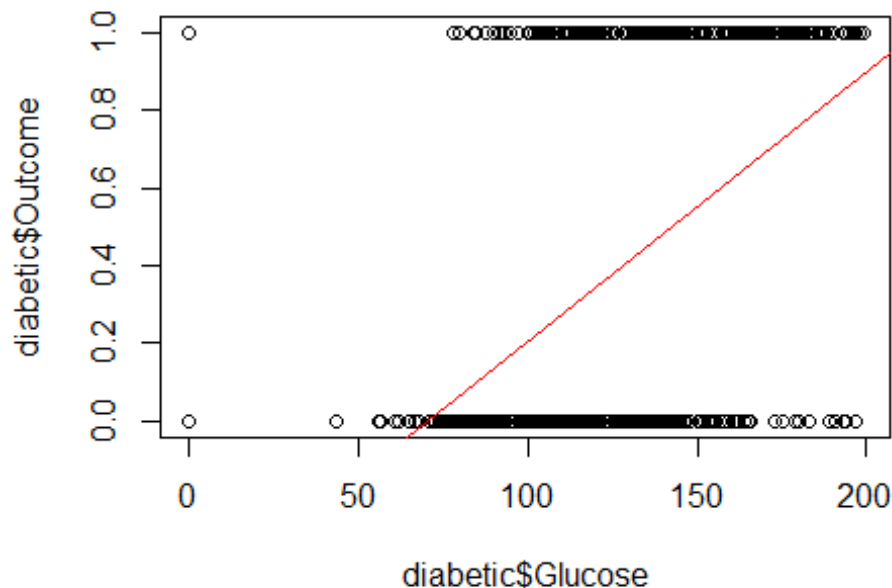
Rahma Sweedy

2 September 2019

```r
diabetic<- read.csv(file.choose(),header=T)
attach(diabetic)
cor(diabetic$Outcome,diabetic$Glucose)
```

```
## [1] 0.4665814
```

```r
plot(diabetic$Outcome~diabetic$Glucose)
model1<- lm(diabetic$Outcome~diabetic$Glucose,data = diabetic)
abline(model1,col="red")
```



```r
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.5.3
```

```r
par(mfrow=c(1, 2))
plot(density(diabetic$Glucose),main = "glucose levels",ylab = "frequencies")
polygon(density(diabetic$Glucose ),col="green")
plot(density(diabetic$Outcome),main = "diseased state",ylab = "frequencies")
polygon(density(diabetic$Outcome ),col="green")
```

glucose levels

N = 768   Bandwidth = 7.337

diseased state

N = 768   Bandwidth = 0.1137
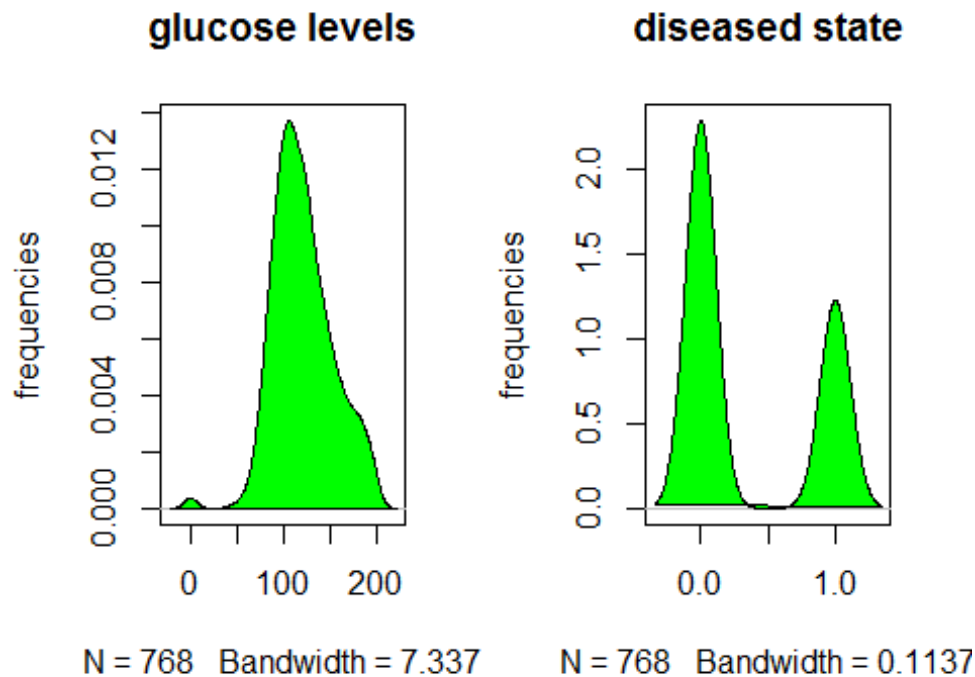
```
summary(model1)

##
## Call:
## lm(formula = diabetic$Outcome ~ diabetic$Glucose, data = diabetic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8787 -0.3027 -0.1479  0.3649  1.4925
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.4924949  0.0596125  -8.262  6.3e-16 ***
## diabetic$Glucose  0.0069602  0.0004767  14.600  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4221 on 766 degrees of freedom
## Multiple R-squared:  0.2177, Adjusted R-squared:  0.2167
## F-statistic: 213.2 on 1 and 766 DF,  p-value: < 2.2e-16
```
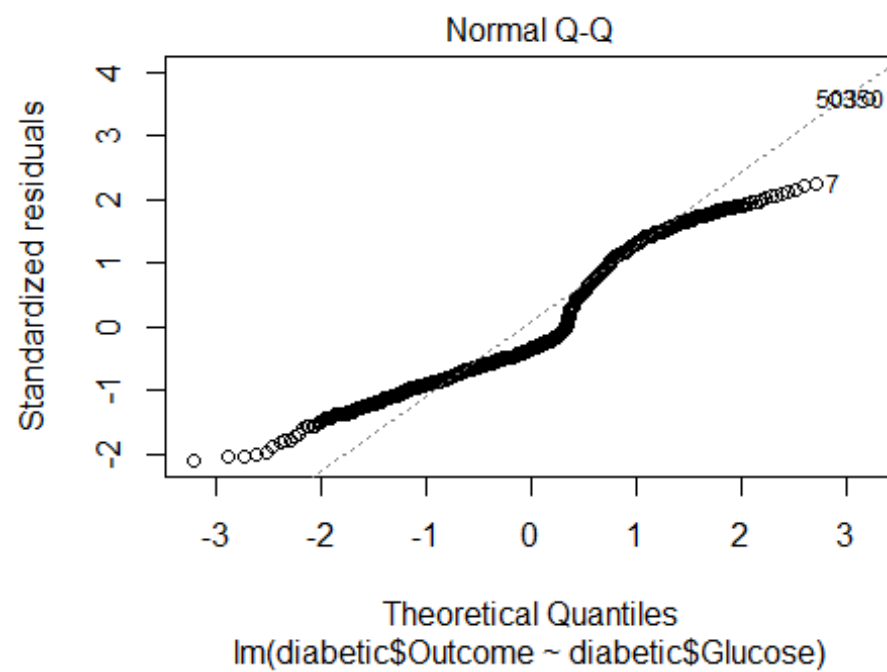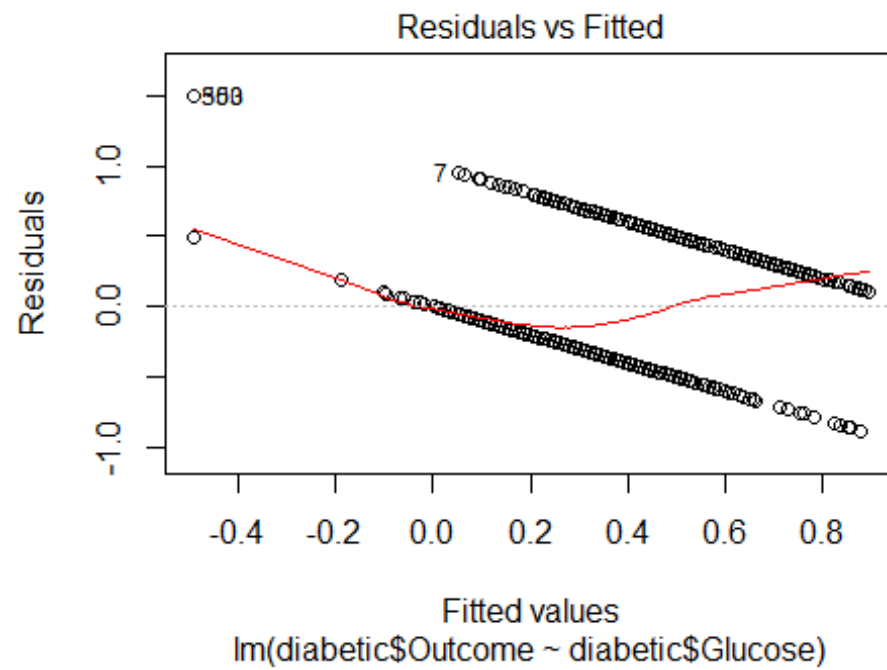
regression eq.
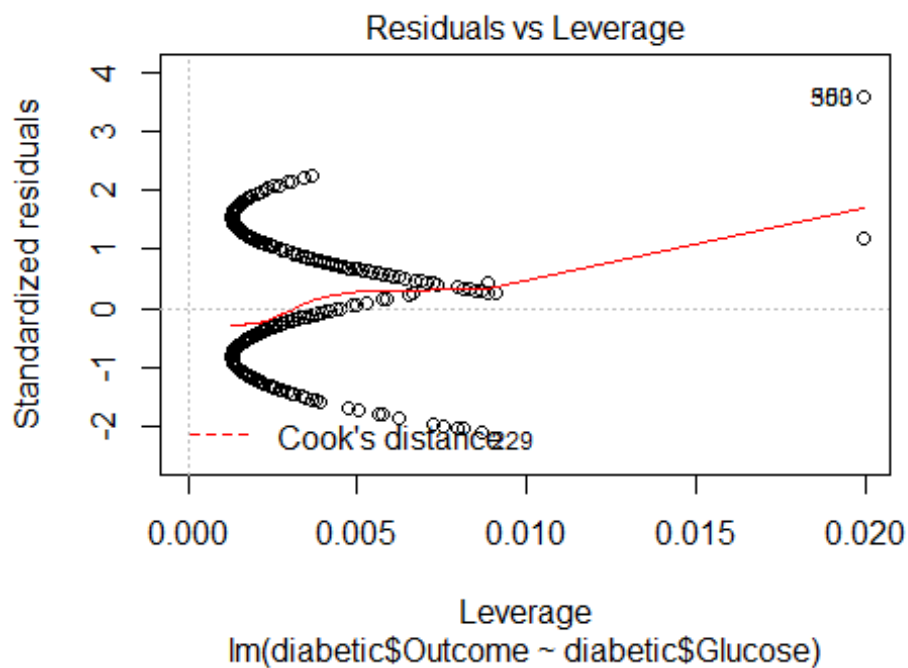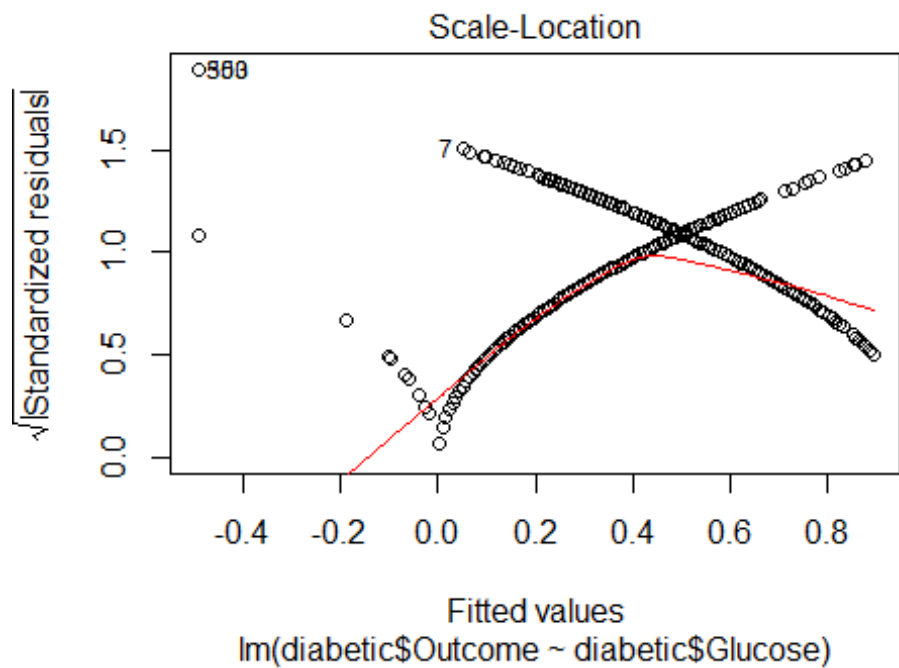
Outcome = (-0.49)+(0.00696*Glucose)

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: diabetic$Outcome
##                     Df  Sum Sq Mean Sq F value    Pr(>F)
## diabetic$Glucose    1   37.984  37.984  213.16 < 2.2e-16 ***
## Residuals          766 136.495   0.178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(model1)
```

Residuals vs Fitted

Im(diabetic$Outcome ~ diabetic$Glucose)

Normal Q-Q

Im(diabetic$Outcome ~ diabetic$Glucose)

Scale-Location

√|Standardized residuals|

Fitted values
lm(diabetic$Outcome ~ diabetic$Glucose)

Residuals vs Leverage

Standardized residuals

Cook's distance

Leverage
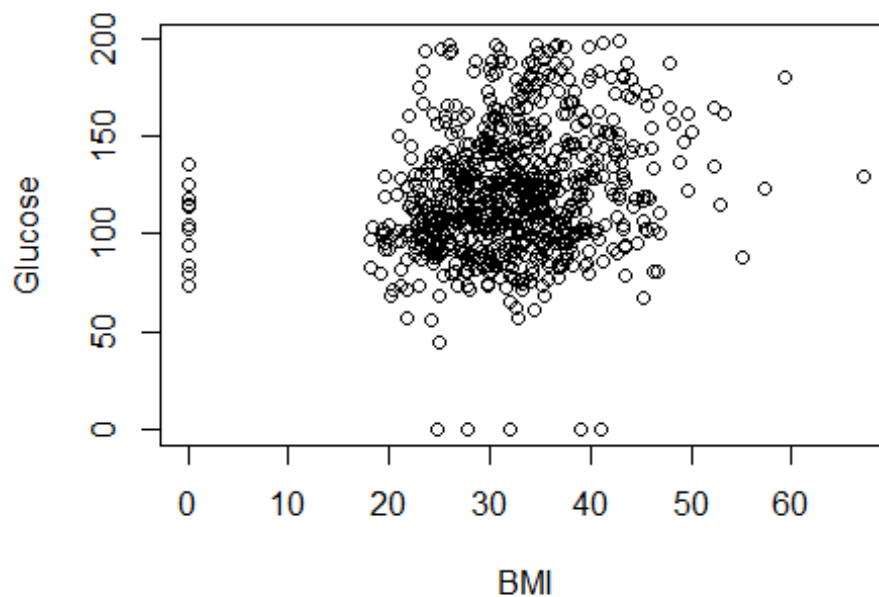lm(diabetic$Outcome ~ diabetic$Glucose)

```
model2<-lm(Glucose~BMI,data = diabetic)
model2
```

```
## 
## Call:
## lm(formula = Glucose ~ BMI, data = diabetic)
## 
## Coefficients:
## (Intercept)          BMI
##     92.2129       0.8965
```
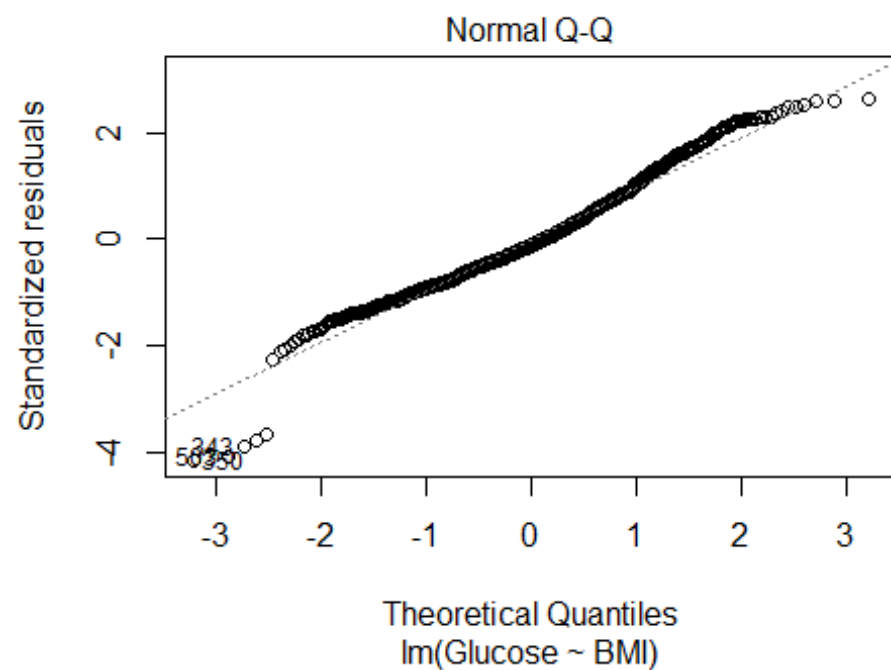
```
cor(Glucose,BMI)
```

```
## [1] 0.2210711
```

```
plot(Glucose~BMI)
```
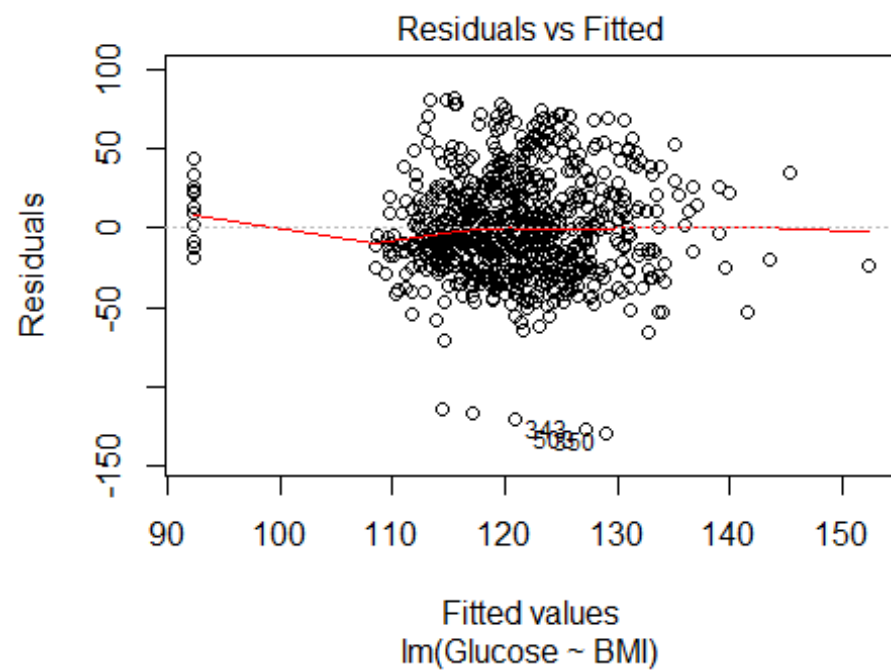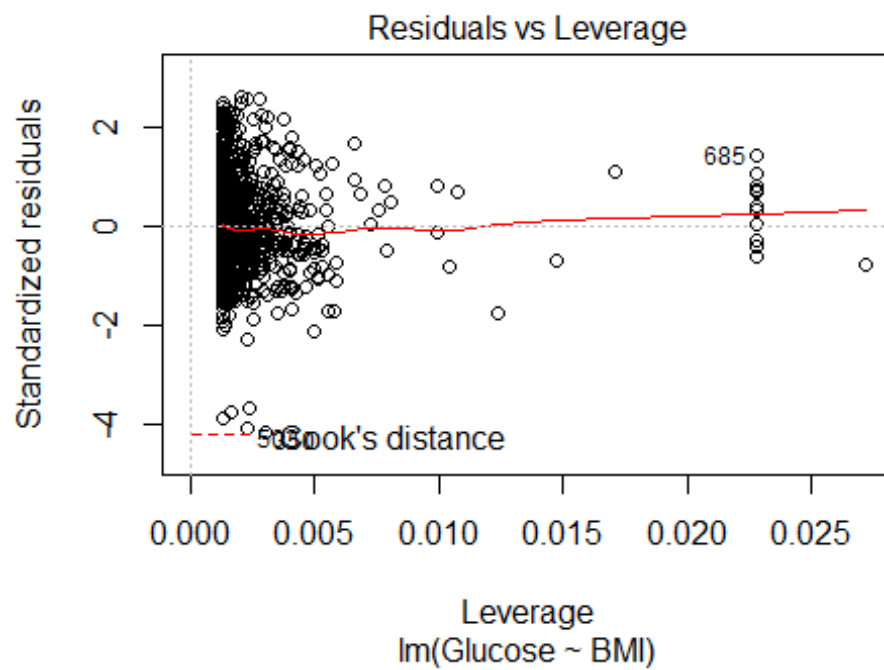


```
summary(model2)
```

```
## 
## Call:
## lm(formula = Glucose ~ BMI, data = diabetic)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -128.970  -20.756   -3.705   19.509   81.568
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  92.2129     4.7083  19.585  < 2e-16 ***
## BMI           0.8965     0.1429   6.274 5.89e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.2 on 766 degrees of freedom
## Multiple R-squared:  0.04887,    Adjusted R-squared:  0.04763
## F-statistic: 39.36 on 1 and 766 DF,  p-value: 5.891e-10
```

```
plot(model2)
```

Residuals vs Fitted

Fitted values
lm(Glucose ~ BMI)

Normal Q-Q

Theoretical Quantiles
lm(Glucose ~ BMI)

Scale-Location

lm(Glucose ~ BMI)

Residuals vs Leverage
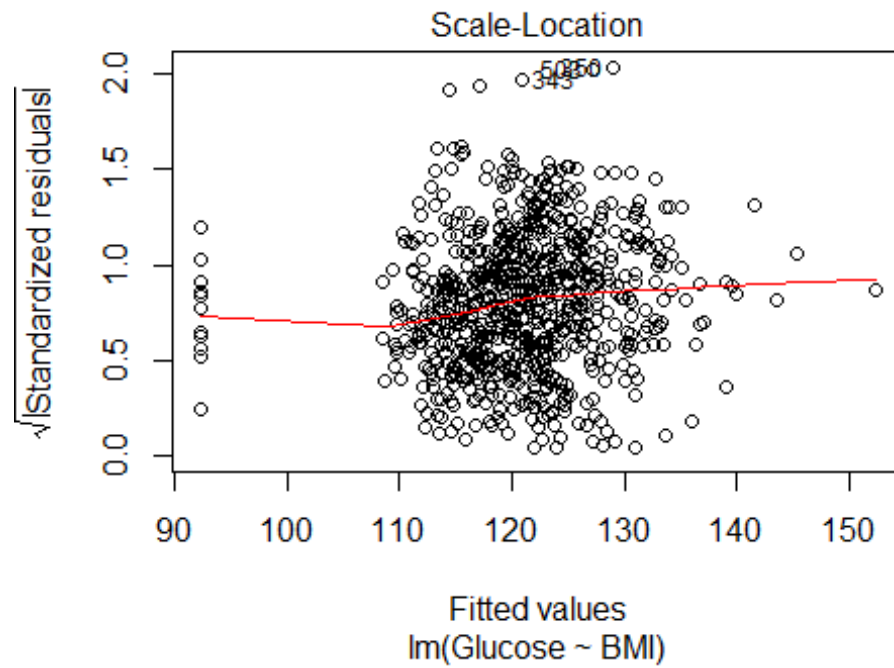
lm(Glucose ~ BMI)

```r
model3<-lm(Glucose~ BMI+Age,data = diabetic)
model3
```

```
## 
## Call:
## lm(formula = Glucose ~ BMI + Age, data = diabetic)
## 
## Coefficients:
## (Intercept)           BMI           Age
##     70.2952        0.8589        0.6955
```

```
summary(model3)
```

```
## 
## Call:
## lm(formula = Glucose ~ BMI + Age, data = diabetic)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -132.310  -19.102   -1.977   18.310   83.766
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.29517    5.40191  13.013  < 2e-16 ***
## BMI          0.85891    0.13808   6.220 8.16e-10 ***
## Age          0.69555    0.09257   7.514 1.61e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 30.13 on 765 degrees of freedom
## Multiple R-squared:  0.1142, Adjusted R-squared:  0.1119
## F-statistic: 49.33 on 2 and 765 DF,  p-value: < 2.2e-16
```

so the age is not confounder

```
model4<- lm(Glucose~BMI+Insulin,data = diabetic)
summary(model4)
```

```
## 
## Call:
## lm(formula = Glucose ~ BMI + Insulin, data = diabetic)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.179  -20.581   -3.339   17.381   86.731
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 93.269854   4.494715  20.751  < 2e-16 ***
## BMI          0.656330   0.139117   4.718 2.83e-06 ***
## Insulin      0.083046   0.009517   8.726  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## 
## Residual standard error: 29.78 on 765 degrees of freedom
## Multiple R-squared:  0.135,  Adjusted R-squared:  0.1327
## F-statistic: 59.68 on 2 and 765 DF,  p-value: < 2.2e-16
```

so insulin is not confounder

```r
model5<-lm(DiabetesPedigreeFunction~SkinThickness,data = diabetic)
summary(model5)
```

```
## 
## Call:
## lm(formula = DiabetesPedigreeFunction ~ SkinThickness, data = diabetic)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.43877 -0.21700 -0.09312  0.15222  1.78659
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3934232  0.0191773  20.515  < 2e-16 ***
## SkinThickness 0.0038202  0.0007377   5.179 2.86e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3259 on 766 degrees of freedom
## Multiple R-squared:  0.03383,    Adjusted R-squared:  0.03257
## F-statistic: 26.82 on 1 and 766 DF,  p-value: 2.856e-07
```

```r
cor(DiabetesPedigreeFunction,SkinThickness)
```

```
## [1] 0.1839276
```

```r
model6<-lm(DiabetesPedigreeFunction~SkinThickness+Age,data = diabetic)
summary(model6)
```

```
## 
## Call:
## lm(formula = DiabetesPedigreeFunction ~ SkinThickness + Age,
##     data = diabetic)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.45570 -0.22281 -0.09691  0.14651  1.79318
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3390035  0.0400631   8.462  < 2e-16 ***
## SkinThickness 0.0039510  0.0007418   5.326 1.32e-07 ***
## Age           0.0015563  0.0010062   1.547    0.122
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3256 on 765 degrees of freedom
## Multiple R-squared:  0.03684,    Adjusted R-squared:  0.03432
## F-statistic: 14.63 on 2 and 765 DF,  p-value: 5.813e-07
```

model7<-**lm**(DiabetesPedigreeFunction~SkinThickness+Insulin,data = diabetic)
summary(model7)

```
##
## Call:
## lm(formula = DiabetesPedigreeFunction ~ SkinThickness + Insulin,
##     data = diabetic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55097 -0.21009 -0.09684  0.14678  1.86025
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3878426  0.0191297  20.274  < 2e-16 ***
## SkinThickness 0.0026460  0.0008148   3.248  0.00121 **
## Insulin       0.0003721  0.0001128   3.299  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3238 on 765 degrees of freedom
## Multiple R-squared:  0.04738,    Adjusted R-squared:  0.04489
## F-statistic: 19.03 on 2 and 765 DF,  p-value: 8.629e-09
```

model8<- **lm**(DiabetesPedigreeFunction~Pregnancies,data = diabetic)
summary(model8)

```
##
## Call:
## lm(formula = DiabetesPedigreeFunction ~ Pregnancies, data = diabetic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4066 -0.2294 -0.1008  0.1511  1.9354
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.484551   0.018149  26.699   <2e-16 ***
## Pregnancies -0.003296   0.003551  -0.928    0.354
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3314 on 766 degrees of freedom
## Multiple R-squared:  0.001124,   Adjusted R-squared:  -0.0001802
## F-statistic: 0.8618 on 1 and 766 DF,  p-value: 0.3535
```
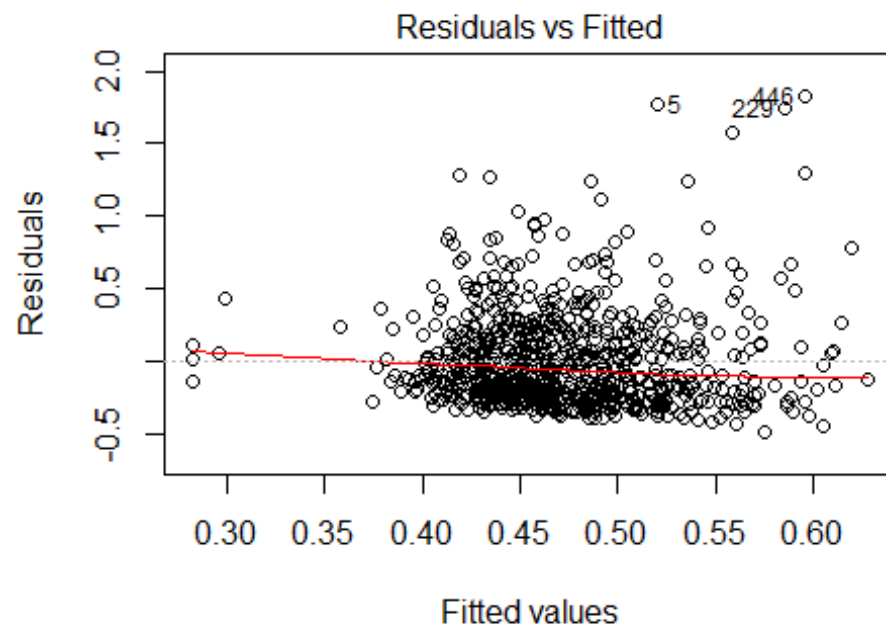
```
model9<-lm(DiabetesPedigreeFunction~Pregnancies+Glucose,data = diabetic)
summary(model9)

##
## Call:
## lm(formula = DiabetesPedigreeFunction ~ Pregnancies + Glucose,
##     data = diabetic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47628 -0.22987 -0.08996  0.15073  1.84014
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3110825  0.0469933   6.620 6.77e-11 ***
## Pregnancies -0.0051305  0.0035465  -1.447    0.148
## Glucose      0.0014932  0.0003738   3.995 7.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3282 on 765 degrees of freedom
## Multiple R-squared:  0.02154,    Adjusted R-squared:  0.01898
## F-statistic:  8.42 on 2 and 765 DF,  p-value: 0.0002416

model10<-
lm(DiabetesPedigreeFunction~Pregnancies+Glucose+(Pregnancies*Glucose),data =
diabetic)
summary(model10)

##
## Call:
## lm(formula = DiabetesPedigreeFunction ~ Pregnancies + Glucose +
##     (Pregnancies * Glucose), data = diabetic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48693 -0.23269 -0.09739  0.15042  1.82428
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.790e-01  7.156e-02   3.899 0.000105 ***
## Pregnancies         3.361e-03  1.471e-02   0.228 0.819359
## Glucose             1.760e-03  5.834e-04   3.016 0.002648 **
## Pregnancies:Glucose -6.845e-05  1.151e-04  -0.595 0.552211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3283 on 764 degrees of freedom
## Multiple R-squared:  0.02199,    Adjusted R-squared:  0.01815
## F-statistic: 5.726 on 3 and 764 DF,  p-value: 0.0007068
```
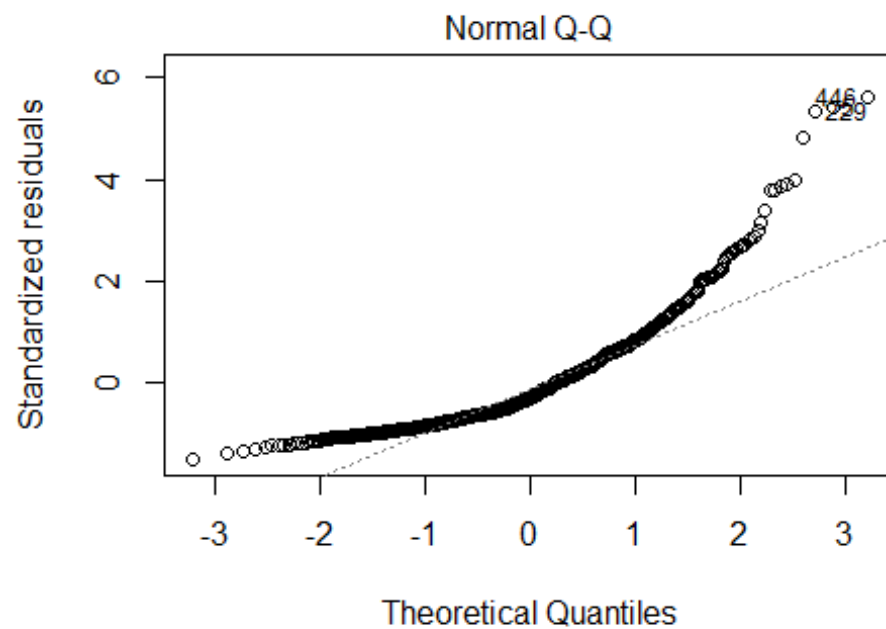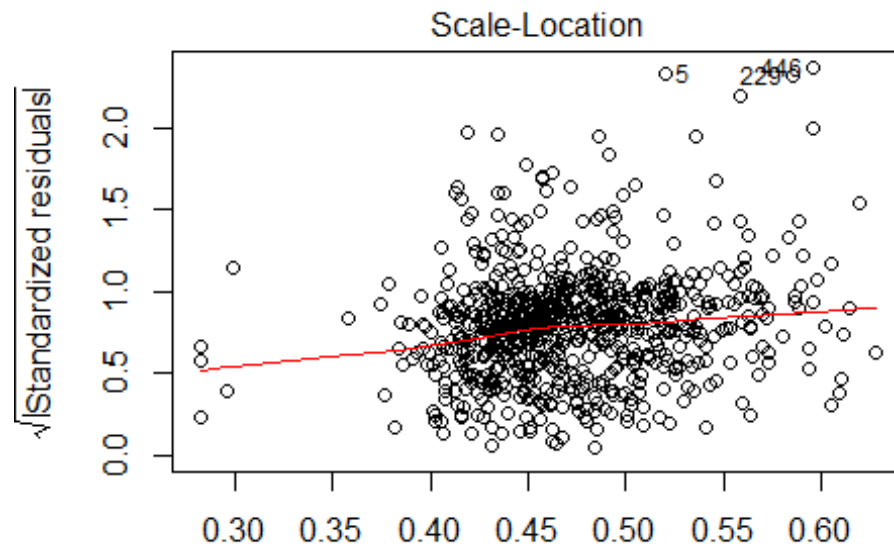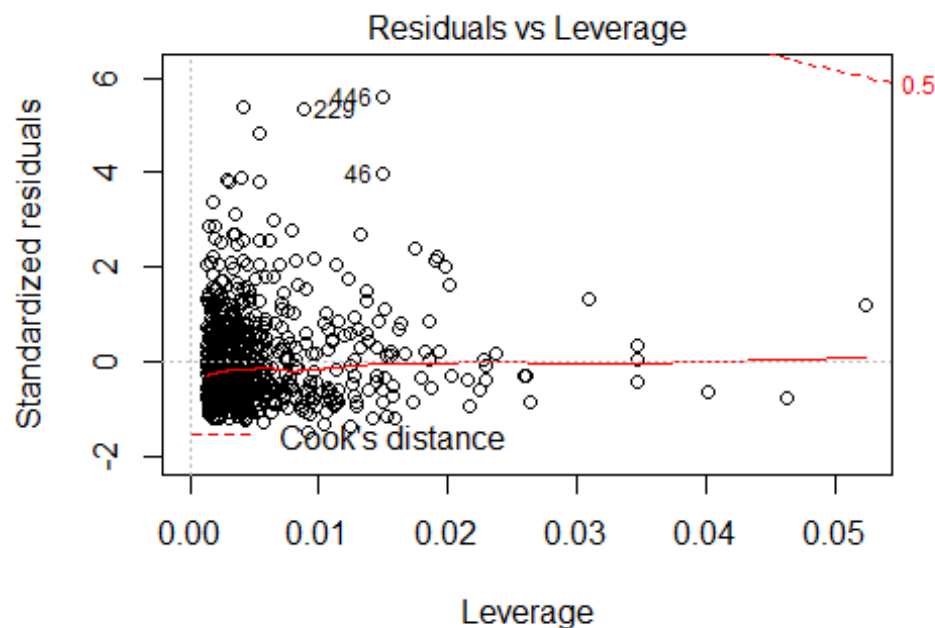
```
par(mfrow=c(1, 1))
plot(model10)
```

Residuals vs Fitted

Fitted values
DiabetesPedigreeFunction ~ Pregnancies + Glucose + (Pregnancies *

Normal Q-Q

Theoretical Quantiles
DiabetesPedigreeFunction ~ Pregnancies + Glucose + (Pregnancies *

Scale-Location

√|Standardized residuals|

5 225 446

Fitted values
DiabetesPedigreeFunction ~ Pregnancies + Glucose + (Pregnancies *

Residuals vs Leverage

Standardized residuals

446 225 46

0.5

Cook's distance

Leverage
DiabetesPedigreeFunction ~ Pregnancies + Glucose + (Pregnancies *