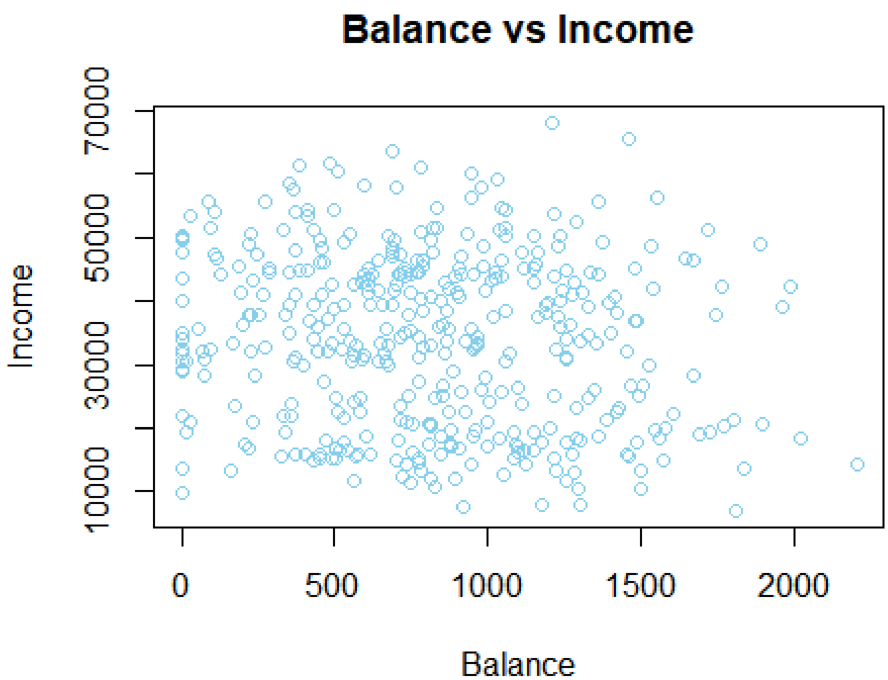
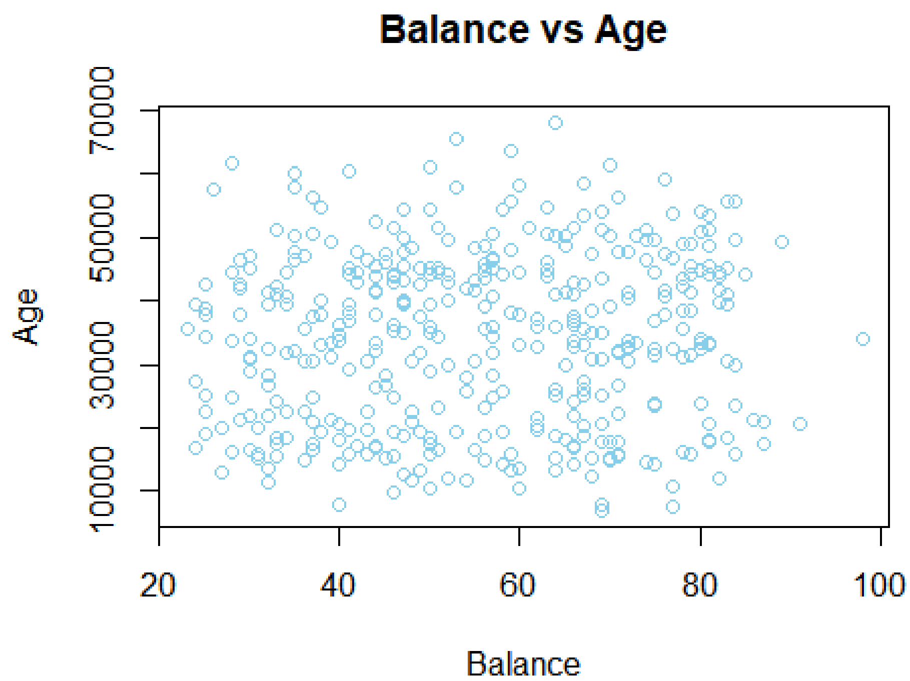
assignment2.R

2019-12-22

```
credit=read.csv("C:/Users/nikhi/Desktop/Credit_Rev-1.csv",header=TRUE)
plot(credit$balance,credit$income,type="p",col="skyblue",xlab="Balance",ylab=
"Income",main="Balance vs Income")
```



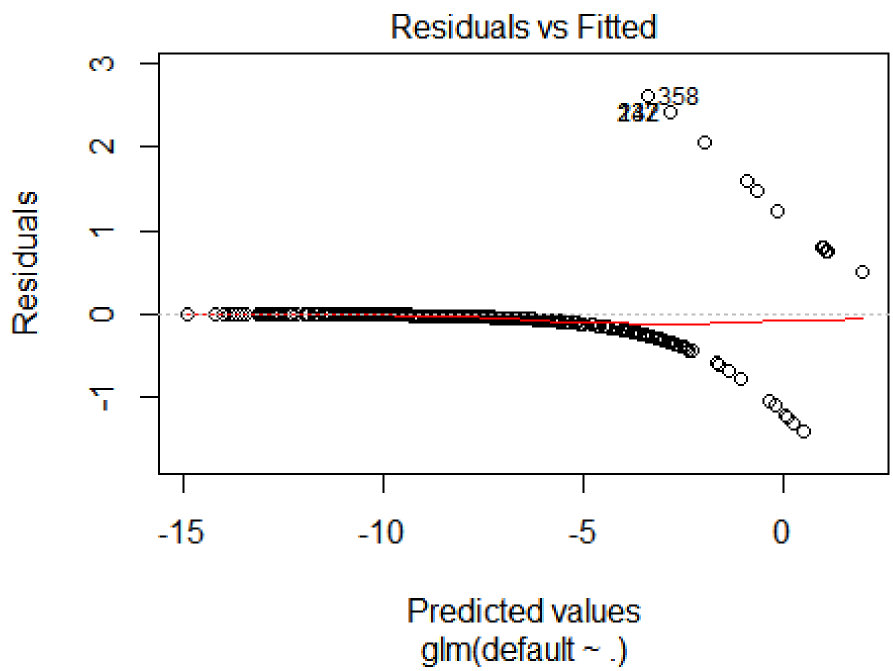

```
plot(credit$Age,credit$income,type="p",col="skyblue",xlab="Balance",ylab="Age
",main="Balance vs Age")
```

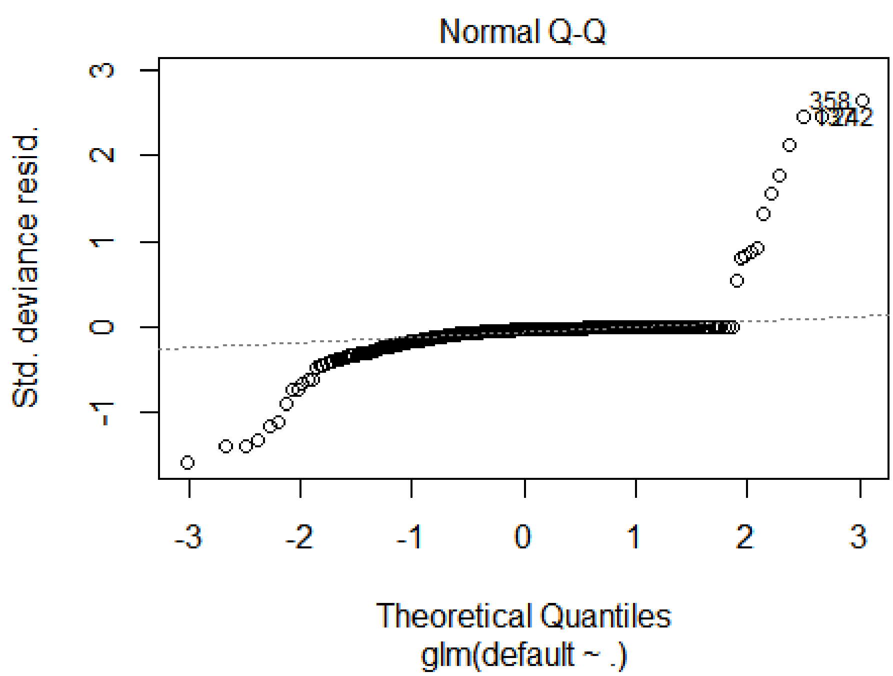


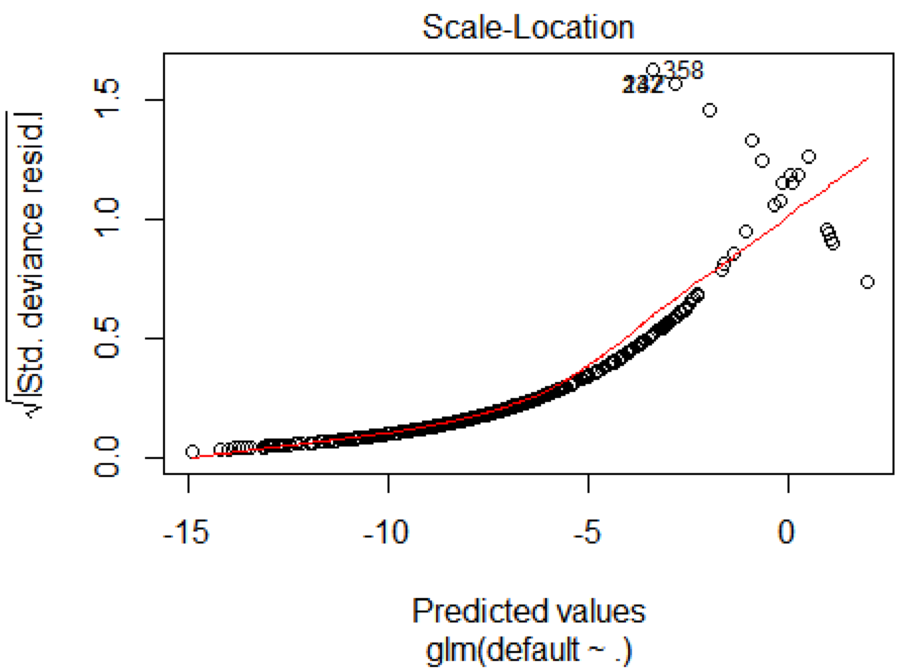

```
cor(credit$balance,credit$income)

## [1] -0.1315461

#Simple logistic regression
glm.fit=glm(default~.,family=binomial,data=credit)
#Default vs student,balance,age,income
#family=binomial to ensure logistic regression is used

print(glm.fit)

##
## Call:  glm(formula = default ~ ., family = binomial, data = credit)
##
## Coefficients:
## (Intercept)            X   studentYes      balance       income          A
ge
##  -1.525e+01    3.900e-03   -6.065e-01    6.988e-03    9.512e-08    2.983e-
02
##
## Degrees of Freedom: 399 Total (i.e. Null);  394 Residual
## Null Deviance:       107.8
## Residual Deviance: 48.19     AIC: 60.19

#Diagnosing model
summary(glm.fit,correlation=TRUE,symbolic.cor=TRUE)
```

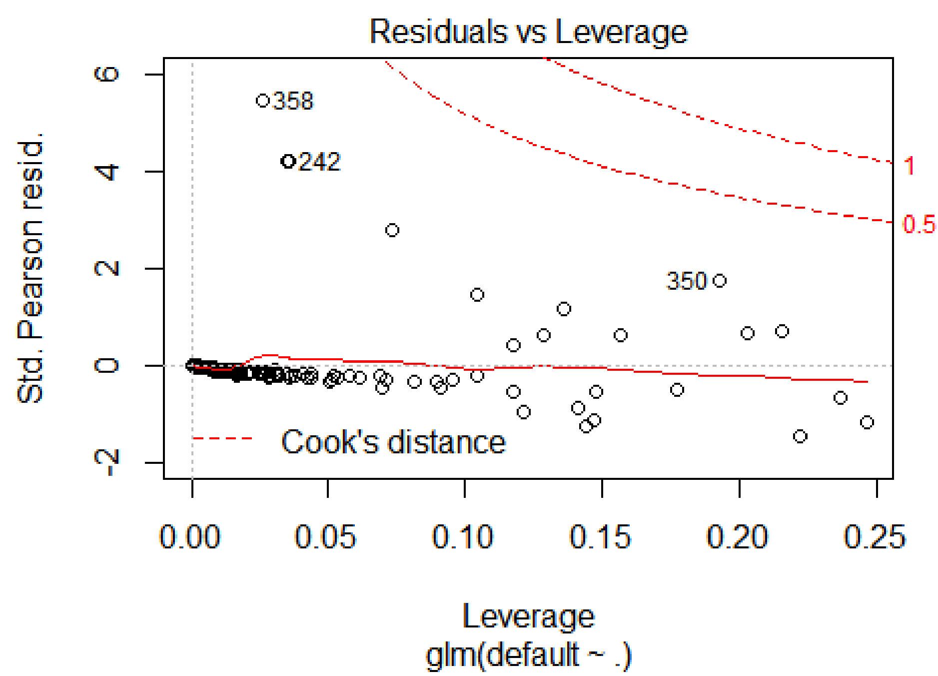
```
##
## Call:
## glm(formula = default ~ ., family = binomial, data = credit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.39715  -0.09518  -0.03107  -0.01097   2.60952
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.525e+01  3.608e+00  -4.227 2.36e-05 ***
## X            3.900e-03  4.254e-03   0.917    0.359
## studentYes  -6.065e-01  1.603e+00  -0.378    0.705
## balance      6.988e-03  1.488e-03   4.697 2.64e-06 ***
## income       9.512e-08  5.456e-05   0.002    0.999
## Age          2.983e-02  2.324e-02   1.284    0.199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 107.794  on 399  degrees of freedom
## Residual deviance:  48.188  on 394  degrees of freedom
## AIC: 60.188
##
## Number of Fisher Scoring iterations: 9
##
## Correlation of Coefficients:
##
## (Intercept) 1
## X           . 1
## studentYes  .   1
## balance     ,     1
## income      .   +   1
## Age         .         1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1

#Null dispersion
#correlation:logical; if TRUE, the correlation matrix of the estimated parame
ters is returned and printed.
#symbolic.cor the correlations in a symbolic form (see symnum) rather than as
numbers.
#the matrix of coefficients, standard errors, z-values and p-values.
#If a 95% confidence interval is available for an absolute measure of interve
ntion effect (e.g. SMD, risk difference, rate difference), then the standard
error can be calculated as

#SE = (upper limit - lower limit) / 3.92.
```

```r
#For 90% confidence intervals divide by 3.29 rather than 3.92; for 99% confid
ence intervals divide by 5.15.



#Where exact P values are quoted alongside estimates of intervention effect,
it is possible to estimate standard errors. While all tests of statistical si
gnificance produce P values, different tests use different mathematical appro
aches to obtain a P value. The method here assumes P values have been obtaine
d through a particularly simple approach of dividing the effect estimate by i
ts standard error and comparing the result (denoted Z) with a standard normal
distribution (statisticians often refer to this as a Wald test). Where signif
icance tests have used other mathematical approaches the estimated standard e
rrors may not coincide exactly with the true standard errors.



#The first step is to obtain the Z value corresponding to the reported P valu
e from a table of the standard normal distribution. A standard error may then
be calculated as

#SE = intervention effect estimate / Z.
#The p-value is a probability.  it is the probability that the observed spati
al pattern was created by some random process. When the p-value is very small
, it means it is very unlikely (small probability)<0.05 that the observed spa
tial pattern is the result of random processes, so you can reject the null hy
pothesis.
#Z-scores are standard deviations. If, for example, a tool returns a z-score
of +2.5, you would say that the result is 2.5 standard deviations

glm.fit$coefficients

##   (Intercept)             X    studentYes       balance        income
## -1.525055e+01  3.899859e-03 -6.064693e-01  6.988318e-03  9.511610e-08
##           Age
##  2.983359e-02

plot(glm.fit)
```


Residuals vs Fitted
Residuals
Predicted values
glm(default ~ .)
358


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles
glm(default ~ .)
358






```
#Confidence interval
confint(glm.fit)

## Waiting for profiling to be done...
```

```
##                     2.5 %      97.5 %
## (Intercept) -2.344439e+01 -9.008494238
## X           -4.074247e-03  0.013140623
## studentYes  -3.782120e+00  2.645416596
## balance      4.491393e-03  0.010461846
## income      -1.083516e-04  0.000108478
## Age         -1.417493e-02  0.078978500
```

```
#Multi logistic regression
#Null dispersion
#correlation:logical; if TRUE, the correlation matrix of the estimated parame
ters is returned and printed.
#symbolic.cor the correlations in a symbolic form (see symnum) rather than as
numbers.
#the matrix of coefficients, standard errors, z-values and p-values.
glm.fit1=glm(default~student+balance+Age+income,family=binomial,data=credit)
print(glm.fit1)
```

```
##
## Call:  glm(formula = default ~ student + balance + Age + income, family =
binomial,
##     data = credit)
##
## Coefficients:
## (Intercept)   studentYes      balance          Age       income
##  -1.449e+01   -8.757e-01    7.261e-03    2.958e-02   -3.872e-06
##
## Degrees of Freedom: 399 Total (i.e. Null);  395 Residual
## Null Deviance:       107.8
## Residual Deviance: 49.08     AIC: 59.08
```

```
summary(glm.fit1,correlation=TRUE,symbolic.cor=TRUE)
```

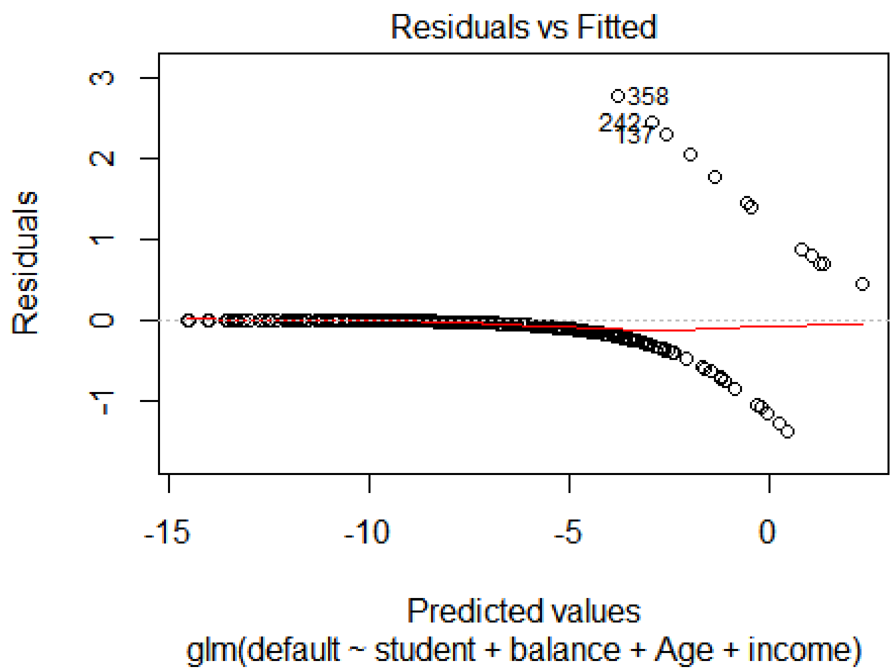
```
##
## Call:
## glm(formula = default ~ student + balance + Age + income, family = binomia
l,
##     data = credit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.37538  -0.10072  -0.03005  -0.01009   2.76116
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.449e+01  3.446e+00  -4.205 2.61e-05 ***
## studentYes  -8.757e-01  1.583e+00  -0.553    0.580
## balance      7.261e-03  1.525e-03   4.763 1.91e-06 ***
## Age          2.958e-02  2.268e-02   1.304    0.192
## income      -3.872e-06  5.427e-05  -0.071    0.943
## ---
```

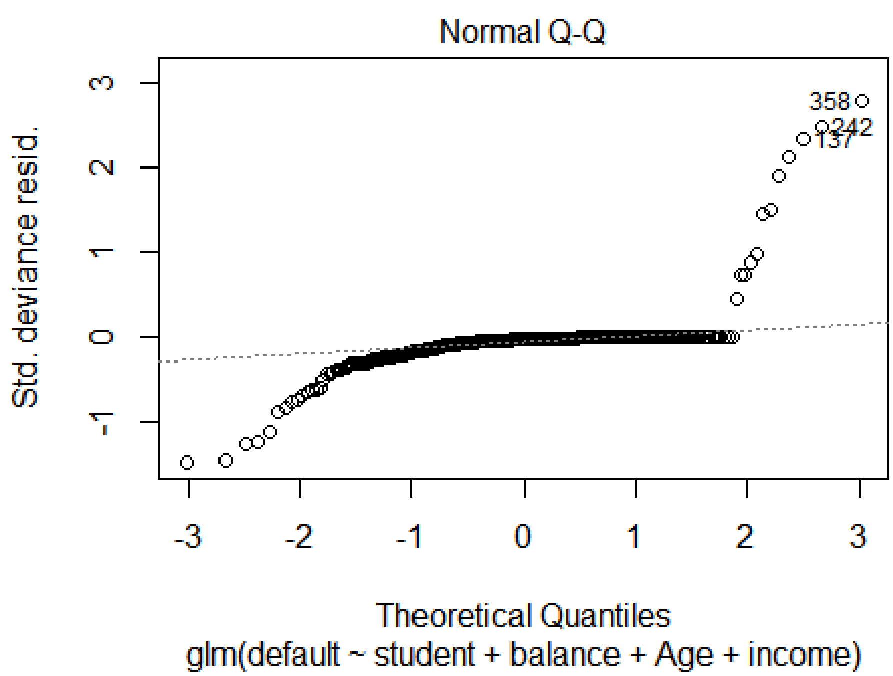
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 107.794  on 399  degrees of freedom
## Residual deviance:  49.079  on 395  degrees of freedom
## AIC: 59.079
##
## Number of Fisher Scoring iterations: 9
##
## Correlation of Coefficients:
##
## (Intercept) 1
## studentYes  . 1
## balance     ,   1
## Age         .     1
## income      . +     1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

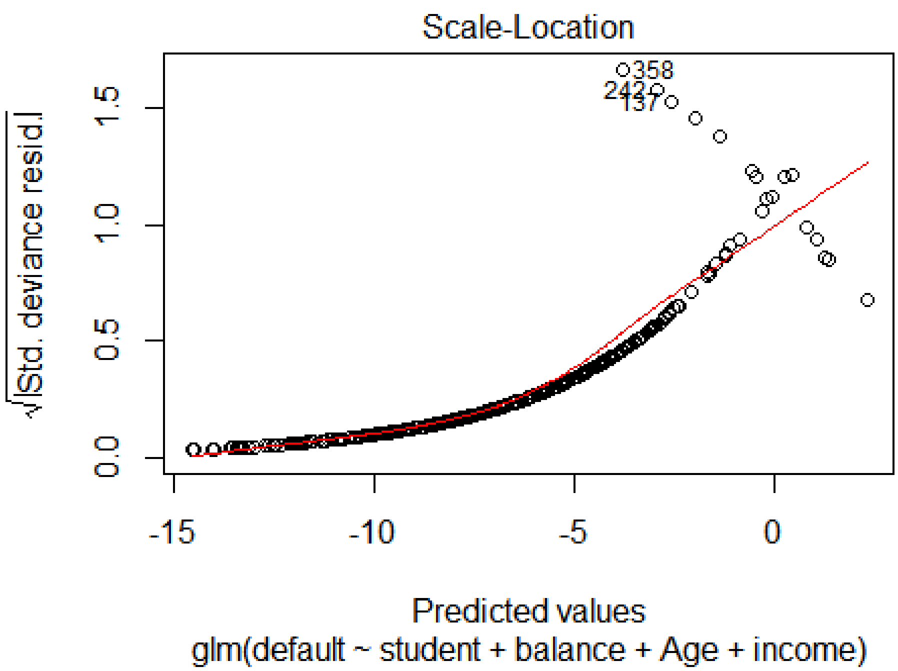
```
glm.fit1$coefficients
```

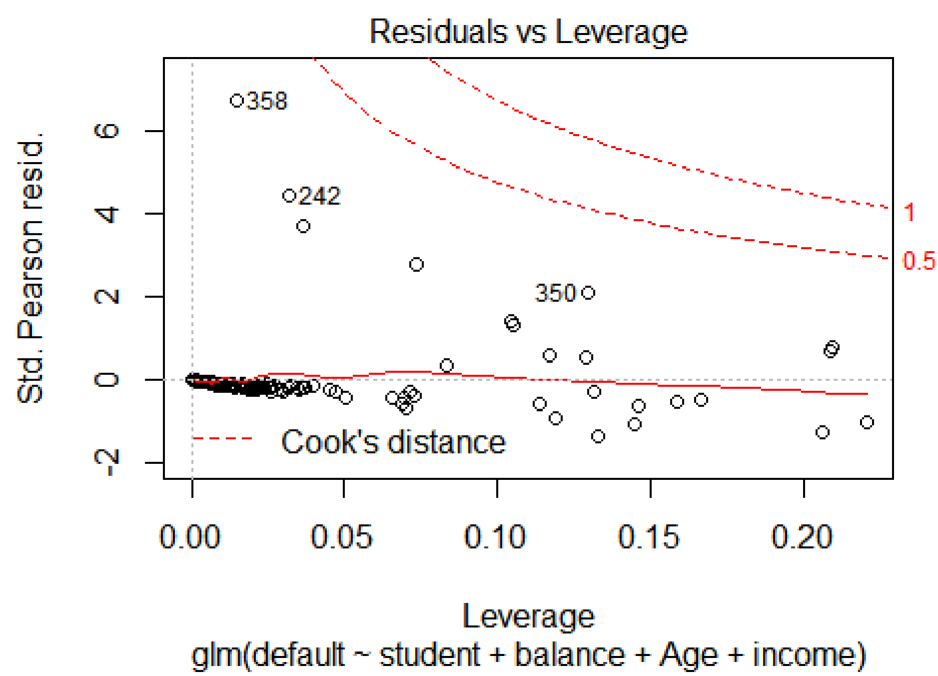
```
##   (Intercept)    studentYes       balance           Age        income
## -1.448768e+01 -8.757040e-01  7.261028e-03  2.958270e-02 -3.872307e-06
```

```
plot(glm.fit1)
```


Residuals vs Fitted
Residuals
Predicted values
glm(default ~ student + balance + Age + income)
358
242
137
-15
-10
-5
0
-1
0
1
2
3


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles
glm(default ~ student + balance + Age + income)
358
242
137
-3
-2
-1
0
1
2
3






```
#Confidence interval
confint(glm.fit1)

## Waiting for profiling to be done...
```

```
##                     2.5 %        97.5 %
## (Intercept) -2.233669e+01 -8.5306070850
## studentYes  -4.018991e+00  2.2989873754
## balance      4.695110e-03  0.0107954769
## Age         -1.326772e-02  0.0774993209
## income      -1.127011e-04  0.0001020003
```

```
X=c(401,402,403,404,405,406)
default=""[-1]
student=c("No","Yes","Yes","No","No","No")
balance=c(1500,1000,2000,2500,1600,1900)
Age=c(34,82,71,36,68,77)
income=c(10000,18000,21000,37000,40000,24000)
newx=c(X,default,student,balance,Age,income)
new=as.data.frame(newx)
pred1=predict(glm.fit,new,type="response")
```

```
## Warning: 'newdata' had 30 rows but variables found have 6 rows
```

```
pred1
```

```
##           1           2           3           4           5           6
## 0.100748646 0.007745337 0.859536411 0.992410108 0.387624889 0.871026879
```

```
pred2=predict(glm.fit1,new,type="response")
```

```
## Warning: 'newdata' had 30 rows but variables found have 6 rows
```

```
pred2
```

```
##           1           2           3           4           5           6
## 0.067292084 0.003184481 0.764538495 0.989915216 0.266333715 0.816558896
```

```
#Probabilities of students defaulting
#Based on the model,students 3,4,5,6 may default while 1,2 will not default
```