SINGAPORE MANAGEMENT UNIVERSITY

DSA211 - Statistical Learning With R

Project Part 1 Report

Section Number: G1

Instructor: Professor Goh Jing Rong

Toh Jing Lin Cheryl	01430750	cheryl.toh.2021
Foo Chuan Wei	01394670	cwfoo.2021
Nguyen Hanh Trang	01422031	htnguyen.2021
Seah Li Ping Megan	01395272	megan.seah.2019
Sharafinaz Binte Shawal	01443439	sharafinazs.2021

1.0 Introduction

In this project, we are assigned to construct and recommend the best multiple regression model to explain the account balance with the given independent variables with reference to the Bank2023P.csv data set.

Our team's general approach is based on the following factors:

1) Achieving a parsimonious model by eliminating redundant independent variables.
2) Elimination of any collinearity issues or model assumption failure
3) Application of nonlinear transformation on independent variables if necessary
4) Exhaustive search of the best model with interaction terms as consideration

2.0 Exploratory Data Analysis - Pair Plots

We first take a closer look at the pair plots (Appendix, 2.0 Output) between the different variables. Upon closer examination, we observed that there is a non-linear relationship, seemingly quadratic, between Income and Balance variables as shown in Figure 1. Additionally, from Figure 2, there appears to be a linear relationship between Limit and Rating.

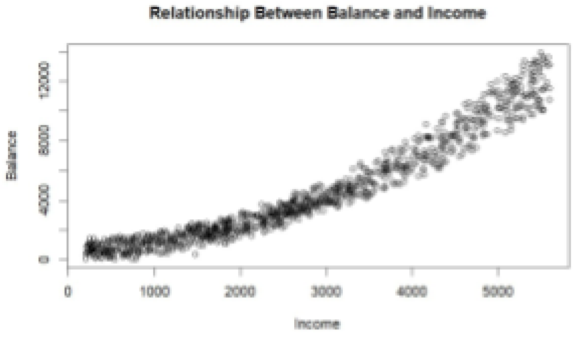


Figure 1: Scatter plot of Balance against Income.

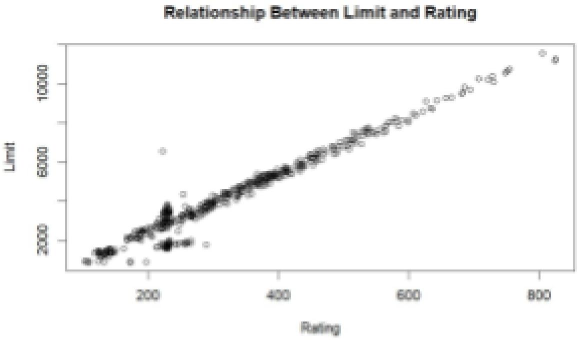


Figure 2: Scatter plot of Limit against Rating.

2.1 Exploratory Data Analysis - Box plots for Categorical Variables

As per our output (Appendix, 2.1 Output), we further examine by utilising box plots of the categorical variables, such as Gender*,* Married and Ethnicity. Based on the box plots, there seems to be no striking difference for the response variable Balance against Male and Female, MarriedYes and MarriedNo as well as Asian, Caucasian and African American.

3.0 Initial Model

We start off with an initial model with all the independent variables and test the model assumptions to determine the validity of using multiple regression for this problem. The equation of the initial model is as follows:

Initial Model: Balance = β_0 + β_1Income + β_2Limit + β_3Rating + β_4Cards + β_5Age + β_6Education + β_7Gender + β_8Married + β_9Ethnicity

The BIC value for the initial model is 16727.15 (Appendix, 3.0 Output).

3.1 Assumptions of Regression

The multiple regression assumptions are listed below:

(1) There is a linear relationship between each independent variable and the dependent variable.

(2) There is no multicollinearity among the independent variables.

(3) Errors are independent.

(4) Errors have a constant variance.

(5) Errors are normally distributed.

Assumptions 1 and 4 are tested using a scatter plot of residuals against the predicted Balance and a scatter plot of residuals against numerical independent variables. Assumptions 1 and 4 are satisfied if the scatter plots show a linear relationship and have no cone-shaped pattern respectively. Since the dataset is not time-series data, we assume that Assumption 3 is satisfied. As for Assumption 2, it is satisfied if the model's Variance Inflationary Factor (VIF) is less than 5. For Assumption 5, it is met if the Kolmogorov-Smirnov (KS) statistic is less than the KS critical value and if the Q-Q and P-P plots follow a diagonal straight line.

3.2 Testing Assumptions for Initial Model

3.2.1 Assumptions 1 and 4

From the scatter plot of the residuals against the predicted values of Balance (Appendix, 3.2.1 Output), a curvilinear relationship was observed, which indicates the presence of non-linear relationship between one or more independent variables and Balance. Hence, Assumption 1 is not satisfied. Along with the scatter plot generated between the Balance and Income variables (Appendix, 3.2.1 Output), we found a nonlinear transformation on Income is necessary (refer to 4.2). Additionally, we also plotted the residuals against all the numerical independent variables

(Appendix, 3.2.1 Output). For all the residual plots, we found that the variance for residuals remains generally constant with higher values of the predicted Balance and the independent variables. Therefore, we conclude that Assumption 4 is satisfied.

3.2.2 Assumption 2

We may interpret the general variance inflation factor (GVIF) by using the same rule of thumb as VIF when squaring values of GVIF^(1/(2*Df)) column. For degree of freedom of 1, we can observe the GVIF values of Limit and Rating to be greater than 5 (Appendix, 3.2.2 Output), implying that Limit and Rating are highly correlated. Assumption 2 is not satisfied. To overcome this issue, we attempted to remove one of the two variables (refer to 4.1).

3.2.3 Assumption 5

We conduct the Kolmogorov-Smirnovm test at the 5% significance level.

H_0 : The residuals are normally distributed

H_a : The residuals are not normally distributed

Based on the KS test results shown in Appendix, 3.2.3 Output, the KS statistic is 0.06880335 > KS critical value of 0.0430098 at the 5% significance level. Hence, we reject the null hypothesis and conclude that there is sufficient evidence to show that the residuals are not normally distributed. However, from the Q-Q and P-P plots (Appendix, 3.2.3 Output), we find that the points in the Q-Q and P-P plots generally follow a diagonal straight line, suggesting that the residuals are normally distributed. Since the KS test is sensitive with large sample sizes (i.e., the KS test concludes significant deviation from normality even with slight deviations from normality), we will conclude based on the Q-Q and P-P plots that Assumption 5 has been satisfied.

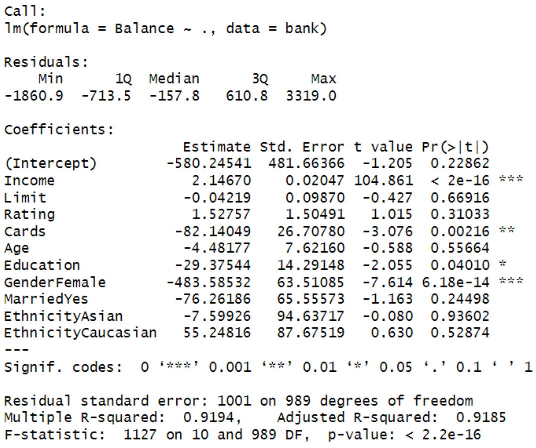
4.0 Elimination of Independent Categorical Variables

$H_0 : \beta_n=0$

$H_a : \beta_n \neq 0$

Where n = (1, no. of coefficients)

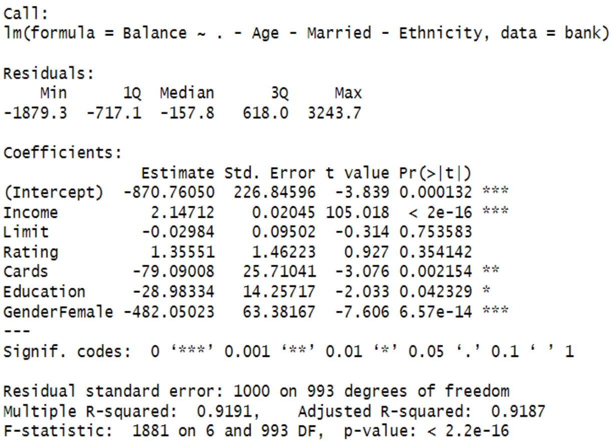
We conduct the above hypothesis test to determine whether a variable is significant at the 5% significance level and eliminate variables if they are not statistically significant.



```
Call:
lm(formula = Balance ~ ., data = bank)

Residuals:
    Min      1Q  Median      3Q     Max 
-1860.9  -713.5  -157.8   610.8  3319.0 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -580.24541  481.66366  -1.205  0.22862    
Income                2.14670    0.02047 104.861  < 2e-16 ***
Limit                -0.04219    0.09870  -0.427  0.66916    
Rating                1.52757    1.50491   1.015  0.31033    
Cards               -82.14049   26.70780  -3.076  0.00216 ** 
Age                  -4.48177    7.62160  -0.588  0.55664    
Education           -29.37544   14.29148  -2.055  0.04010 *  
GenderFemale       -483.58532   63.51085  -7.614 6.18e-14 ***
MarriedYes          -76.26186   65.55573  -1.163  0.24498    
EthnicityAsian       -7.59926   94.63717  -0.080  0.93602    
EthnicityCaucasian   55.24816   87.67519   0.630  0.52874    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1001 on 989 degrees of freedom
Multiple R-squared:  0.9194,	Adjusted R-squared:  0.9185 
F-statistic:  1127 on 10 and 989 DF,  p-value: < 2.2e-16
```

Figure 3: Initial model regression

BIC = 16727.15



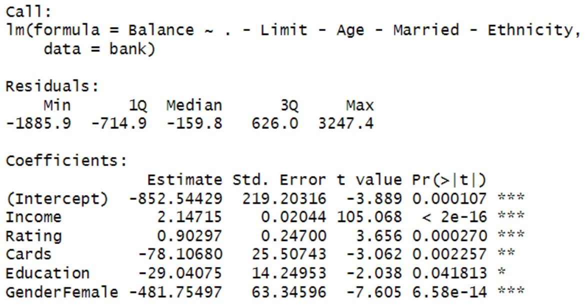
```
Call:
lm(formula = Balance ~ . - Age - Married - Ethnicity, data = bank)

Residuals:
    Min      1Q  Median      3Q     Max 
-1879.3  -717.1  -157.8   618.0  3243.7 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -870.76050  226.84596  -3.839 0.000132 ***
Income          2.14712    0.02045 105.018  < 2e-16 ***
Limit          -0.02984    0.09502  -0.314 0.753583    
Rating          1.35551    1.46223   0.927 0.354142    
Cards         -79.09008   25.71041  -3.076 0.002154 ** 
Education     -28.98334   14.25717  -2.033 0.042329 *  
GenderFemale -482.05023   63.38167  -7.606 6.57e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1000 on 993 degrees of freedom
Multiple R-squared:  0.9191,	Adjusted R-squared:  0.9187 
F-statistic:  1881 on 6 and 993 DF,  p-value: < 2.2e-16
```

Figure 4: Regression model without age, married and ethnicity

BIC = 16702.1

By conducting hypothesis testing on the initial model as shown in Figure 3, the p-value for Married, Ethnicity and Age variables are more than 0.05. Additionally, the box plots and pair plots (Appendix, 2.0 and 2.1 Output) do not show much correlation for these three categorical variables. Hence, we fail to reject the null hypothesis and conclude that these 3 independent categorical variables are not significant.

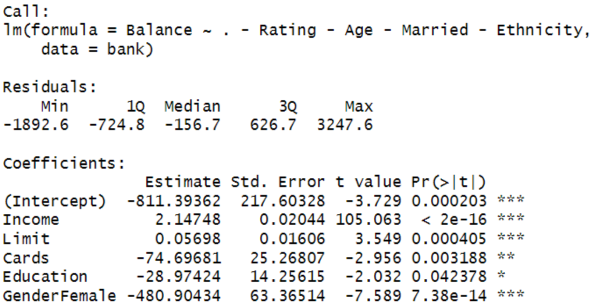
4.1 Linear Relationship Between Limit and Rating Variable



```
Call:
lm(formula = Balance ~ . - Limit - Age - Married - Ethnicity, 
    data = bank)

Residuals:
    Min      1Q  Median      3Q     Max 
-1885.9  -714.9  -159.8   626.0  3247.4 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -852.54429  219.20316  -3.889 0.000107 ***
Income          2.14715    0.02044 105.068  < 2e-16 ***
Rating          0.90297    0.24700   3.656 0.000270 ***
Cards         -78.10680   25.50743  -3.062 0.002257 ** 
Education     -29.04075   14.24953  -2.038 0.041813 *  
GenderFemale -481.75497   63.34596  -7.605 6.58e-14 ***
```

Figure 5: Regression model without Limit variable

BIC = 16695.29



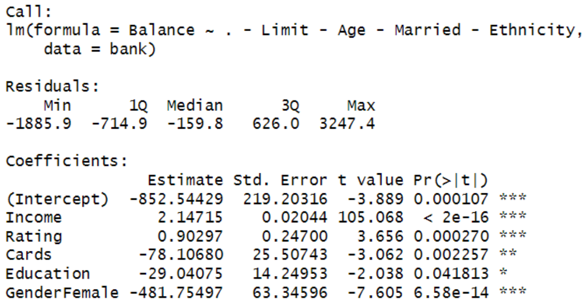
```
Call:
lm(formula = Balance ~ . - Rating - Age - Married - Ethnicity, 
    data = bank)

Residuals:
    Min      1Q  Median      3Q     Max 
-1892.6  -724.8  -156.7   626.7  3247.6 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -811.39362  217.60328  -3.729 0.000203 ***
Income          2.14748    0.02044 105.063  < 2e-16 ***
Limit           0.05698    0.01606   3.549 0.000405 ***
Cards         -74.69681   25.26807  -2.956 0.003188 ** 
Education     -28.97424   14.25615  -2.032 0.042378 *  
GenderFemale -480.90434   63.36514  -7.589 7.38e-14 ***
```

Figure 6: Regression model without Rating variable

BIC = 16696.06

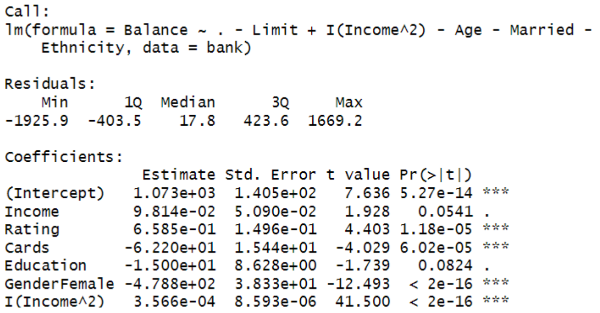
Since the pairplot (Appendix, 2.0 Output) shows a linear relationship between Limit and Rating and hence a collinear issue, we have to eliminate one of the 2 variables and compare the BIC values. From Figures 5 and 6, we observed that once we remove one of 2 variables, the other variable becomes statistically significant at the 5% level. Overall, the BIC value has reduced. However, there are minimal differences in BIC when removing either Limit or Rating. We take the final regression with the lowest BIC by removing the Limit variable.

4.2 Quadratic Relationship Between Balance and Income Variable



```
Call:
lm(formula = Balance ~ . - Limit - Age - Married - Ethnicity,
    data = bank)

Residuals:
    Min      1Q  Median      3Q     Max
-1885.9  -714.9  -159.8   626.0  3247.4

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -852.54429  219.20316  -3.889 0.000107 ***
Income         2.14715    0.02044 105.068  < 2e-16 ***
Rating         0.90297    0.24700   3.656 0.000270 ***
Cards        -78.10680   25.50743  -3.062 0.002257 **
Education    -29.04075   14.24953  -2.038 0.041813 *
GenderFemale -481.75497   63.34596  -7.605 6.58e-14 ***
```

Figure 7: Regression model without Income^2 variable



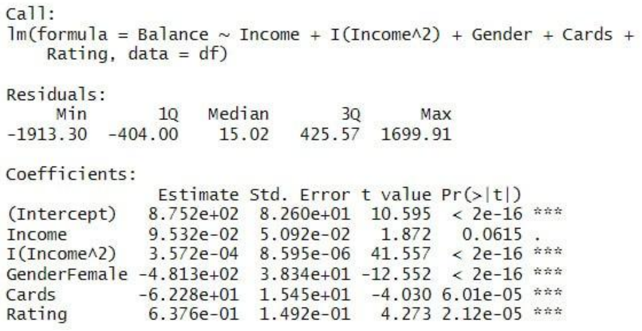
```
Call:
lm(formula = Balance ~ . - Limit + I(Income^2) - Age - Married -
    Ethnicity, data = bank)

Residuals:
    Min      1Q  Median      3Q     Max
-1925.9  -403.5    17.8   423.6  1669.2

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.073e+03  1.405e+02   7.636 5.27e-14 ***
Income        9.814e-02  5.090e-02   1.928   0.0541 .
Rating        6.585e-01  1.496e-01   4.403 1.18e-05 ***
Cards        -6.220e+01  1.544e+01  -4.029 6.02e-05 ***
Education    -1.500e+01  8.628e+00  -1.739   0.0824 .
GenderFemale -4.788e+02  3.833e+01 -12.493  < 2e-16 ***
I(Income^2)   3.566e-04  8.593e-06  41.500  < 2e-16 ***
```

Figure 8: Regression model with Income^2 variable

BIC = 16695.29

BIC = 15696.3

Based on the scatter plot generated between Balance and Income variables (Appendix, 2.1 Output), there seemed to be a quadratic relationship between the two variables. Therefore, we attempted to add $Income^2$ as an additional variable. The BIC value has decreased drastically, showing a better regression model. From Figure 8, upon adding the $Income^2$ variable into the regression, we can infer that $Income^2$ is statistically significant at the 5% level, showing that there is indeed presence of a quadratic relationship. In addition, the p-value for Education variable is larger than 0.05. With that, we fail to reject the null hypothesis and Education is deemed not significant and eliminated from the regression model. Hence, we conclude the best model excluding interaction terms to be Model 4. Referring to Figure 9, excluding Income, as the p-values of all the variables in Model 4 are <0.05, it indicates that the variables are all significant. Model 4 would be the best model thus far as the BIC value is lower than that of the initial model.



```
Call:
lm(formula = Balance ~ Income + I(Income^2) + Gender + Cards +
    Rating, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1913.30  -404.00    15.02   425.57  1699.91

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.752e+02  8.260e+01  10.595  < 2e-16 ***
Income        9.532e-02  5.092e-02   1.872   0.0615 .
I(Income^2)   3.572e-04  8.595e-06  41.557  < 2e-16 ***
GenderFemale -4.813e+02  3.834e+01 -12.552  < 2e-16 ***
Cards        -6.228e+01  1.545e+01  -4.030 6.01e-05 ***
Rating        6.376e-01  1.492e-01   4.273 2.12e-05 ***
```

Figure 9: Regression Model without variables

BIC of Model 4 = 15692.43

Model 4: Balance = β_0 + β_1Income + $\beta_2 Income^2$ + β_3GenderFemale + β_4Cards + β_5Rating

4.3 Identifying Interaction Effects

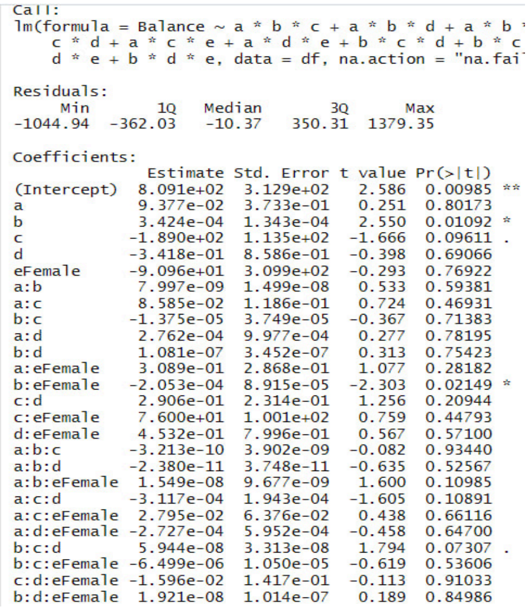
We used the same hypothesis testing as 4.0 to identify whether there are any statistically significant interaction terms. The test shall be done only on the potential interaction terms as including interaction terms may sometimes make the main effect terms less significant. With that, we have the below hypotheses as well at the 5% level of significance:

H_0 : There is no interaction between variables

H_a : There is a significant interaction between variables

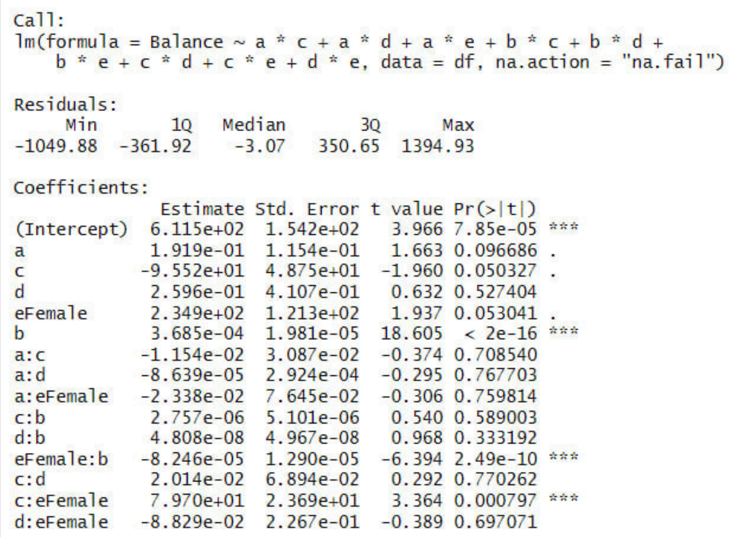
As per 4.2, the predictor variables we will consider interacting with are $Income^2$, Income, Gender, Rating and Cards and we assign them a,b,c,d,e respectively for accountability and readability in the code. We will also consider only up to the 3rd degree of interaction terms to prevent further model complexity.

4.3.1 Model Selection of Interaction effects



```
Call:
lm(formula = Balance ~ a * b * c + a * b * d + a * b
    c * d + a * c * e + a * d * e + b * c * d + b * c
    d * e + b * d * e, data = df, na.action = "na.fai

Residuals:
     Min       1Q   Median       3Q      Max 
-1044.94  -362.03   -10.37   350.31  1379.35 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)   
(Intercept)     8.091e+02  3.129e+02   2.586  0.00985 **
a               9.377e-02  3.733e-01   0.251  0.80173   
b               3.424e-04  1.343e-04   2.550  0.01092 * 
c              -1.890e+02  1.135e+02  -1.666  0.09611 . 
d              -3.418e-01  8.586e-01  -0.398  0.69066   
eFemale        -9.096e+01  3.099e+02  -0.293  0.76922   
a:b             7.997e-09  1.499e-08   0.533  0.59381   
a:c             8.585e-02  1.186e-01   0.724  0.46931   
b:c            -1.375e-05  3.749e-05  -0.367  0.71383   
a:d             2.762e-04  9.977e-04   0.277  0.78195   
b:d             1.081e-07  3.452e-07   0.313  0.75423   
a:eFemale       3.089e-01  2.868e-01   1.077  0.28182   
b:eFemale      -2.053e-04  8.915e-05  -2.303  0.02149 * 
c:d             2.906e-01  2.314e-01   1.256  0.20944   
c:eFemale       7.600e+01  1.001e+02   0.759  0.44793   
d:eFemale       4.532e-01  7.996e-01   0.567  0.57100   
a:b:c          -3.213e-10  3.902e-09  -0.082  0.93440   
a:b:d          -2.380e-11  3.748e-11  -0.635  0.52567   
a:b:eFemale     1.549e-08  9.677e-09   1.600  0.10985   
a:c:d          -3.117e-04  1.943e-04  -1.605  0.10891   
a:c:eFemale     2.795e-02  6.376e-02   0.438  0.66116   
a:d:eFemale    -2.727e-04  5.952e-04  -0.458  0.64700   
b:c:d           5.944e-08  3.313e-08   1.794  0.07307 . 
b:c:eFemale    -6.499e-06  1.050e-05  -0.619  0.53606   
c:d:eFemale    -1.596e-02  1.417e-01  -0.113  0.91033   
b:d:eFemale     1.921e-08  1.014e-07   0.189  0.84986   
```

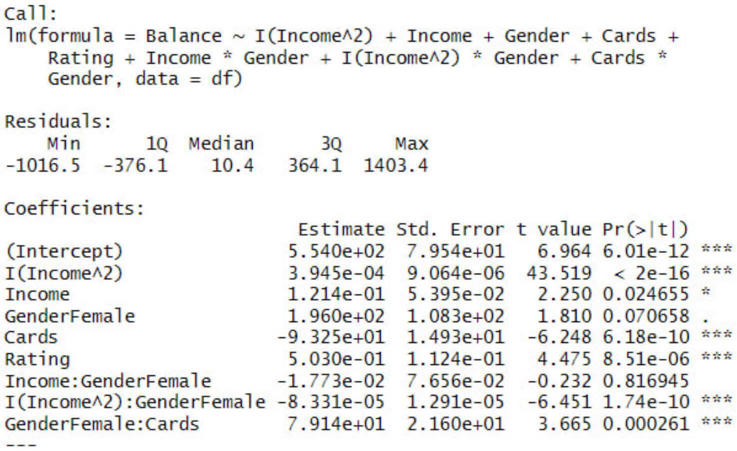
Figure 10: Regression model of 3rd degree interactions



```
Call:
lm(formula = Balance ~ a * c + a * d + a * e + b * c + b * d + 
    b * e + c * d + c * e + d * e, data = df, na.action = "na.fail")

Residuals:
     Min       1Q   Median       3Q      Max 
-1049.88  -361.92    -3.07   350.65  1394.93 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.115e+02  1.542e+02   3.966 7.85e-05 ***
a            1.919e-01  1.154e-01   1.663 0.096686 .  
c           -9.552e+01  4.875e+01  -1.960 0.050327 .  
d            2.596e-01  4.107e-01   0.632 0.527404    
eFemale      2.349e+02  1.213e+02   1.937 0.053041 .  
b            3.685e-04  1.981e-05  18.605  < 2e-16 ***
a:c         -1.154e-02  3.087e-02  -0.374 0.708540    
a:d         -8.639e-05  2.924e-04  -0.295 0.767703    
a:eFemale   -2.338e-02  7.645e-02  -0.306 0.759814    
c:b          2.757e-06  5.101e-06   0.540 0.589003    
d:b          4.808e-08  4.967e-08   0.968 0.333192    
eFemale:b   -8.246e-05  1.290e-05  -6.394 2.49e-10 ***
c:d          2.014e-02  6.894e-02   0.292 0.770262    
c:eFemale    7.970e+01  2.369e+01   3.364 0.000797 ***
d:eFemale   -8.829e-02  2.267e-01  -0.389 0.697071    
```

Figure 11: Regression model of 2nd degree interactions

From our hypothesis testing as per Figure 10, there were no statistically significant 3rd degree interaction terms at the 5% level of significance. With that, we looked at the 2nd degree interactions as per Figure 11. We noted two terms that were statistically significant which were between Income (specifically $Income^2$) *Gender and Cards*Gender whose p-values were less than 0.05 and hence we were able to reject the null hypothesis. The removal of the non-significant interaction terms resulted in this final current regression model as per Figure 12.



```
Call:
lm(formula = Balance ~ I(Income^2) + Income + Gender + Cards +
    Rating + Income * Gender + I(Income^2) * Gender + Cards *
    Gender, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1016.5  -376.1    10.4   364.1  1403.4

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.540e+02  7.954e+01   6.964 6.01e-12 ***
I(Income^2)               3.945e-04  9.064e-06  43.519  < 2e-16 ***
Income                    1.214e-01  5.395e-02   2.250 0.024655 *
GenderFemale              1.960e+02  1.083e+02   1.810 0.070658 .
Cards                    -9.325e+01  1.493e+01  -6.248 6.18e-10 ***
Rating                    5.030e-01  1.124e-01   4.475 8.51e-06 ***
Income:GenderFemale      -1.773e-02  7.656e-02  -0.232 0.816945
I(Income^2):GenderFemale -8.331e-05  1.291e-05  -6.451 1.74e-10 ***
GenderFemale:Cards        7.914e+01  2.160e+01   3.665 0.000261 ***
---
```

Figure 12: Best Regression Model with Interaction terms

BIC Value: 15138.18 (Appendix 4.3.1 Output)

From the p-value output of Figure 12, p-values of $Income^2$*Gender and Cards*Gender is all less than 0.05. We can reject the null hypothesis and conclude that there is sufficient evidence to show there is interaction between these factors. Furthermore, the p-value of the independent variable Rating remained less than 0.05 despite the introduction of interaction terms and hence remained statistically significant as well. Hence, we conclude Model 5 as the most complete model with interaction terms. Overall, we can also conclude that Model 5 is the best model as it has a lower BIC value 15138.18 as compared to Model 4’s 15692.43.

Model 5 : Balance = β_0 + β_1Income + $\beta_2 Income^2$ + β_3GenderFemale + β_4Cards + β5 Rating + β6 Cards*GenderFemale + β_7GenderFemale*Income + $\beta_8 Income^2$*Gender

5.0 Testing Assumptions for Proposed Model

5.1 Assumption 1 and 4

We use residual analysis to test Assumptions 1 and 4. Based on the plot of residuals against predicted Balance in the proposed model (Appendix, 5.1 Output), we can see there is a linear relationship between Balance and residuals. Hence, Assumption 1 is satisfied. Additionally, we also plotted the residuals against all the numerical independent variables (Appendix, 5.1 Output). The residual plots do not show a cone-shape, indicating that the residuals have constant variance. Therefore, we conclude that Assumption 4 is satisfied.

5.2 Assumption 2

From the table (Appendix, 5.2 Output), except for Cards and Rating, the VIFs for the remaining variables are greater than 5, indicating multicollinearity is present among these variables. However, the high degree of multicollinearity could be due to the structural multicollinearity when creating the quadratic variable $Income^2$ from the independent variable Income as well as including the interaction terms. With that, we conclude that the high degree of multicollinearity does not affect the predicted balance. The multicollinearity issue can be ignored and Assumption 2 is satisfied.

5.3 Assumption 5

We conduct the KS test at 5% significance level on the proposed model (Appendix, 5.3 Output) with the following hypothesis:

H_0: The residuals are normally distributed

H_a: The residuals are not normally distributed

Since KS statistic is 0.0496299 > KS critical value of 0.0430098, we can reject the null hypothesis and conclude that there is sufficient evidence to show that the residuals are not normally distributed. However, based on the P-P and Q-Q plot, observations generally follow a diagonal straight line, indicating that the residuals are normally distributed. Since the KS test is sensitive with large sample sizes (i.e., the KS test concludes significant deviation from normality even with slight deviations from normality), we will conclude based on the Q-Q and P-P plots that Assumption 5 has been satisfied.

6.0 Conclusion

With our model assumptions sound, we can use the regression in Model 5 to make inferences about the dependent variable Balance. We conclude that Model 5 is the best model with the lowest BIC value of 15138.18 as compared to Model 4's and initial model as it considered introducing quadratic relationship and interaction terms.

Model 5 : Balance = 554 + 0.1214Income + 3.945e-04$Income^2$ + 196GenderFemale - 93.25Cards + 0.503Rating + 79.14Cards*GenderFemale - 1.773e-02GenderFemale*Income - 8.331e-05GenderFemale*$Income^2$

Appendix: R Inputs and Outputs

All figures used in the report may be found in the appendix

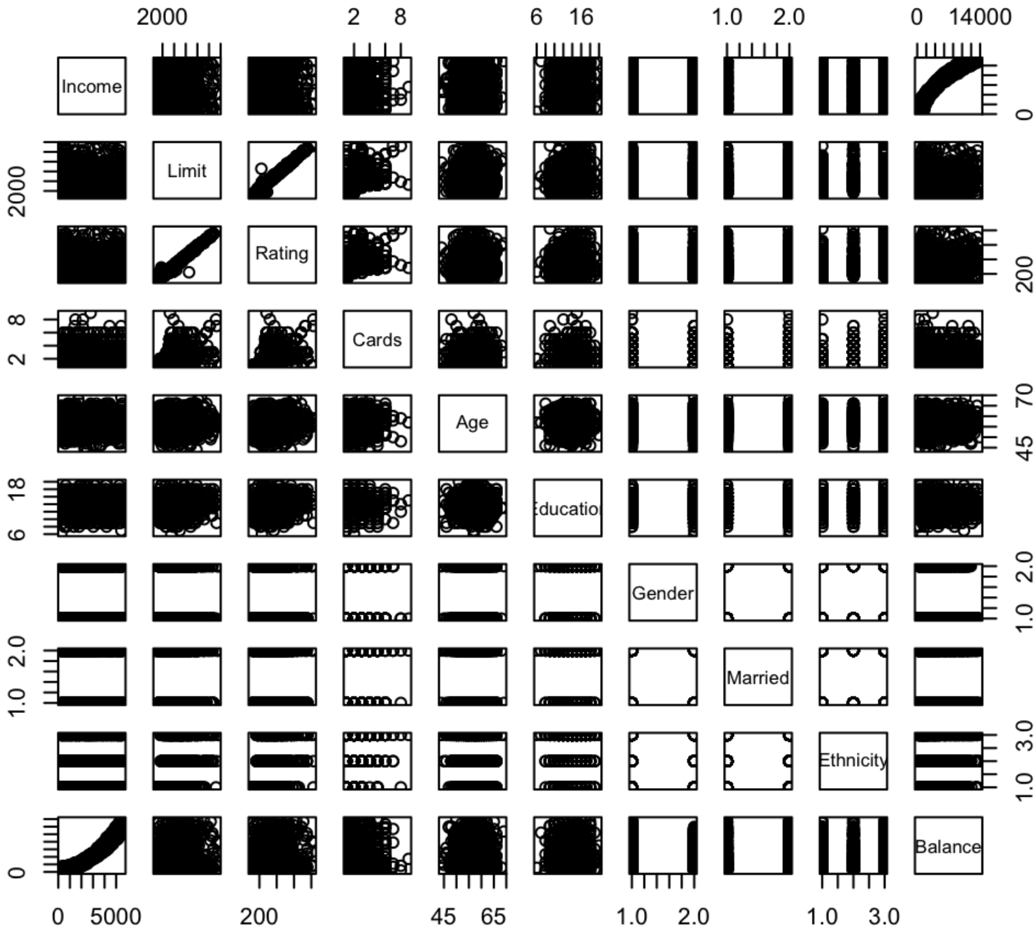
2.0 Input

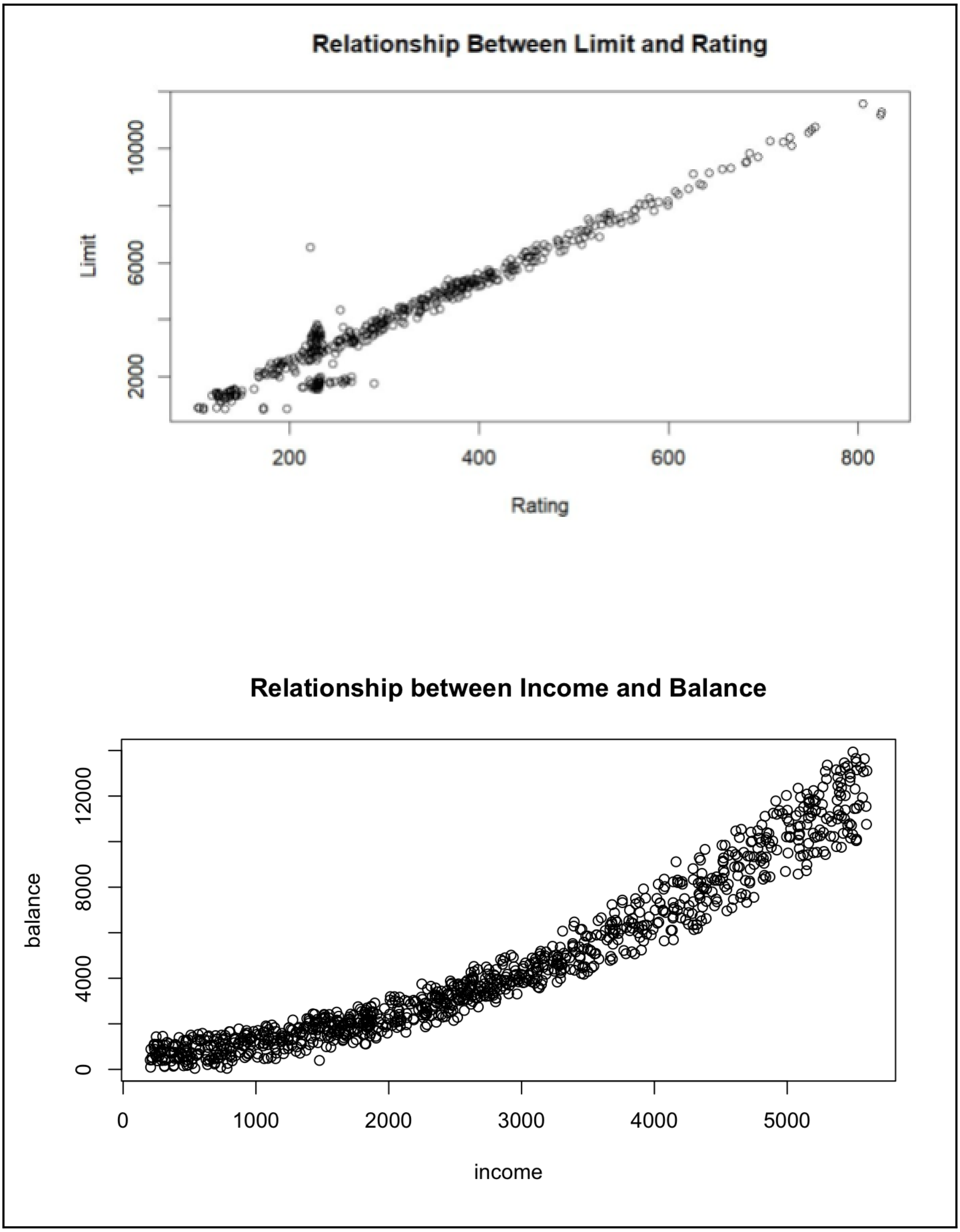
```
df <- read.csv("Bank2023P.csv", stringsAsFactors = TRUE)
attach(df)
#spotting linear relationships
pairs(df)
```

```
plot(Income, Balance, main = "Relationship between Income and Balance",   xlab = "income", ylab = "balance")
```

```
plot(Rating, Limit, main = "Relationship between Limit and Rating", xlab = "Rating", ylab = "Limit")
```

2.0 Output





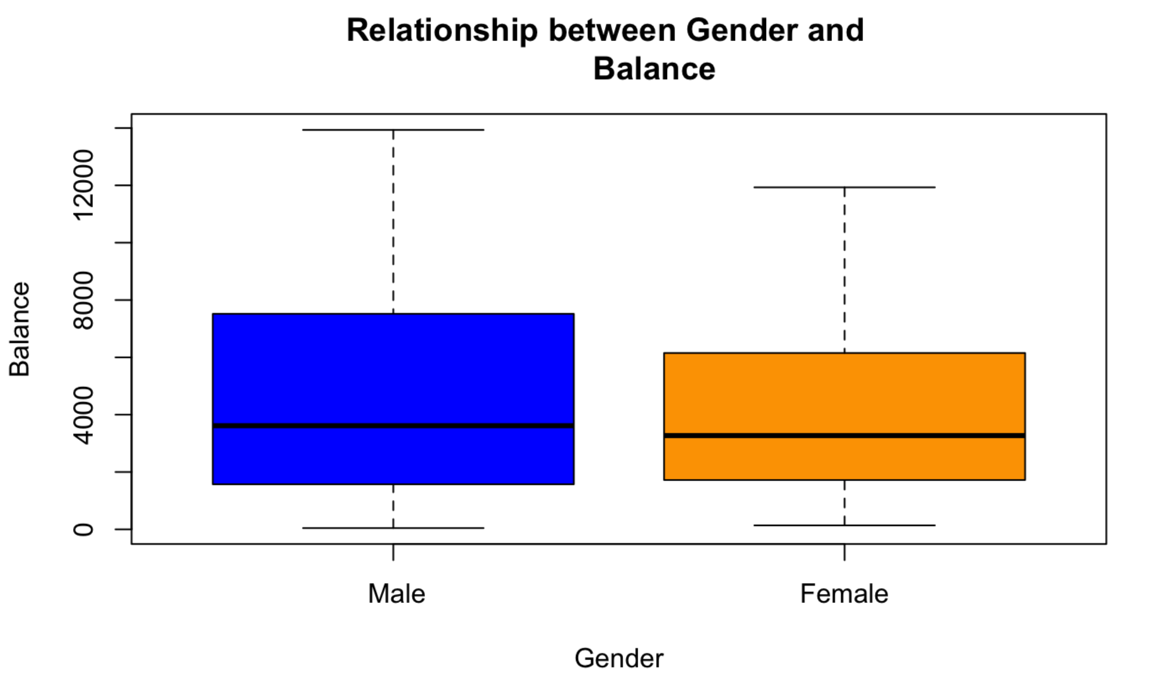
Relationship Between Limit and Rating

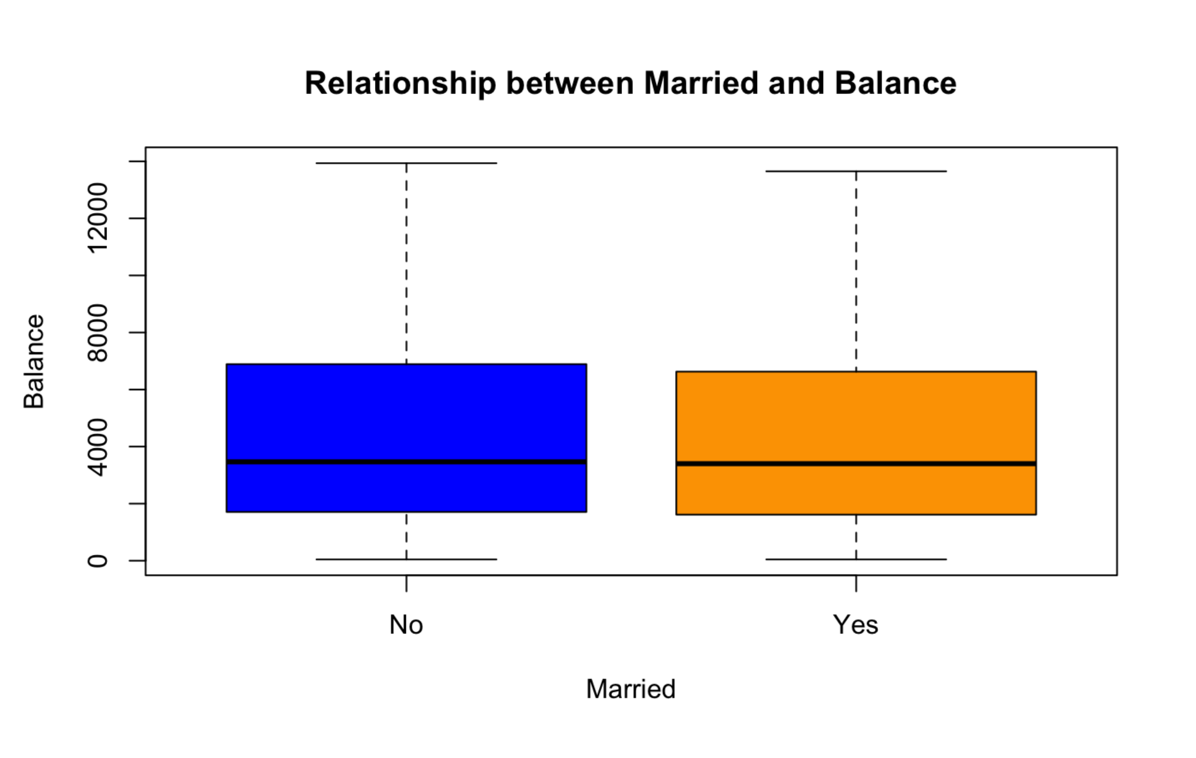
Relationship between Income and Balance

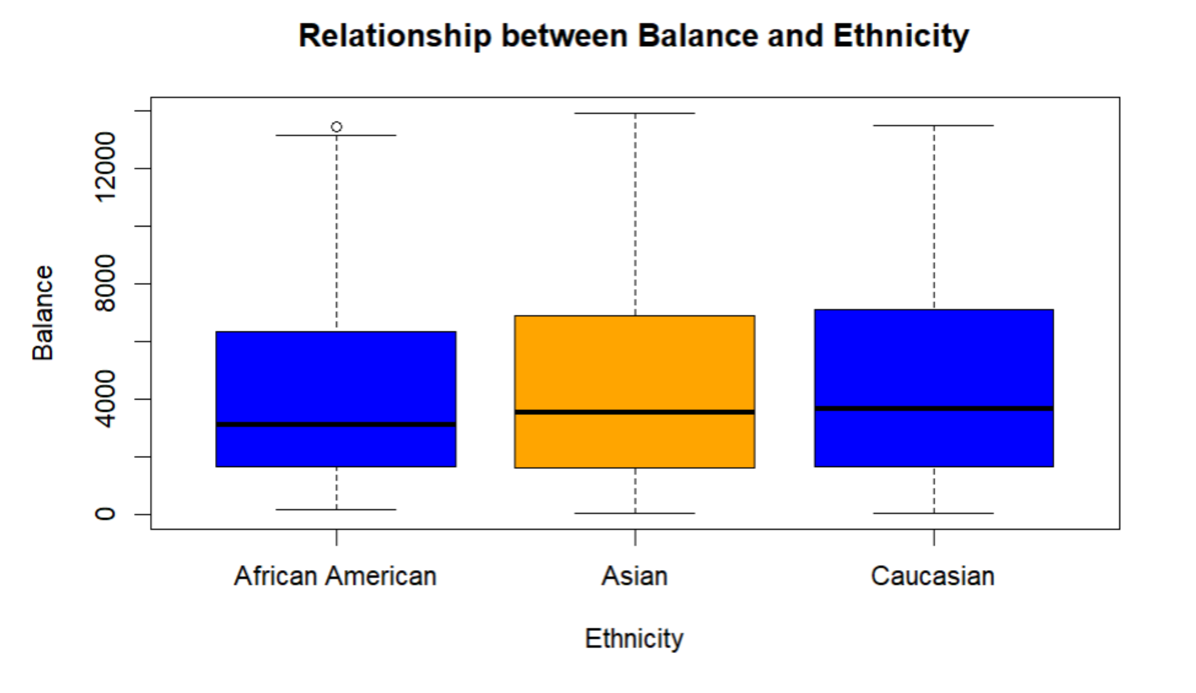
2.1 Input

```
boxplot(Balance~Gender, col=c("blue", "orange"), main = "Relationship between
Gender and Balance", xlab="Gender", ylab="Balance")

boxplot(Balance~Married, col=c("blue", "orange"), main = "Relationship between
Married and Balance", lab="Married", ylab="Balance")

boxplot(Balance~Ethnicity, col=c("blue", "orange"), main = "Relationship between
Ethnicity and Balance", xlab="Ethnicity", ylab="Balance")
```

2.1 Output



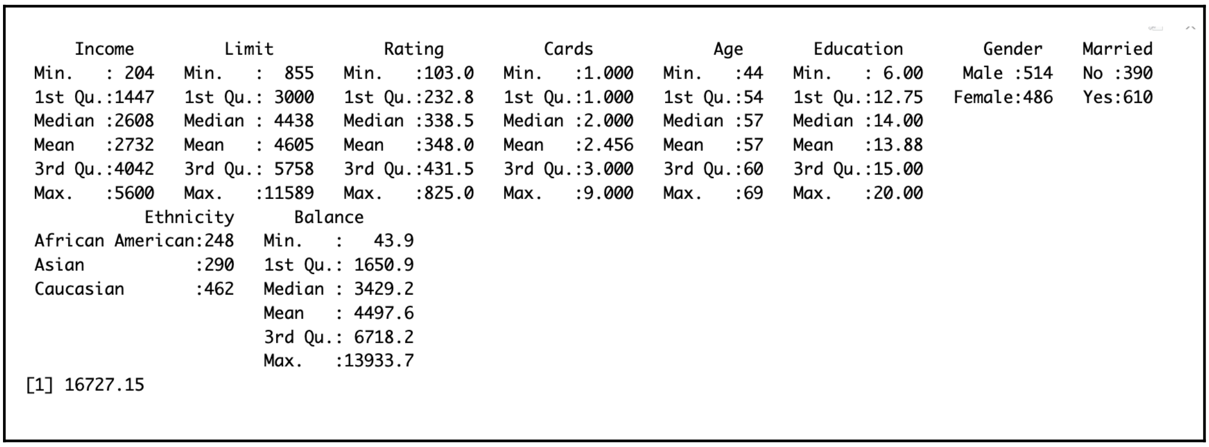




3.0 Input

```
bank <- read.csv("Bank2023P.csv", stringsAsFactors = TRUE)
attach(bank)
summary(bank)
initialModel <- lm(Balance~., data = bank)
BIC(initialModel)
```

3.0 Output

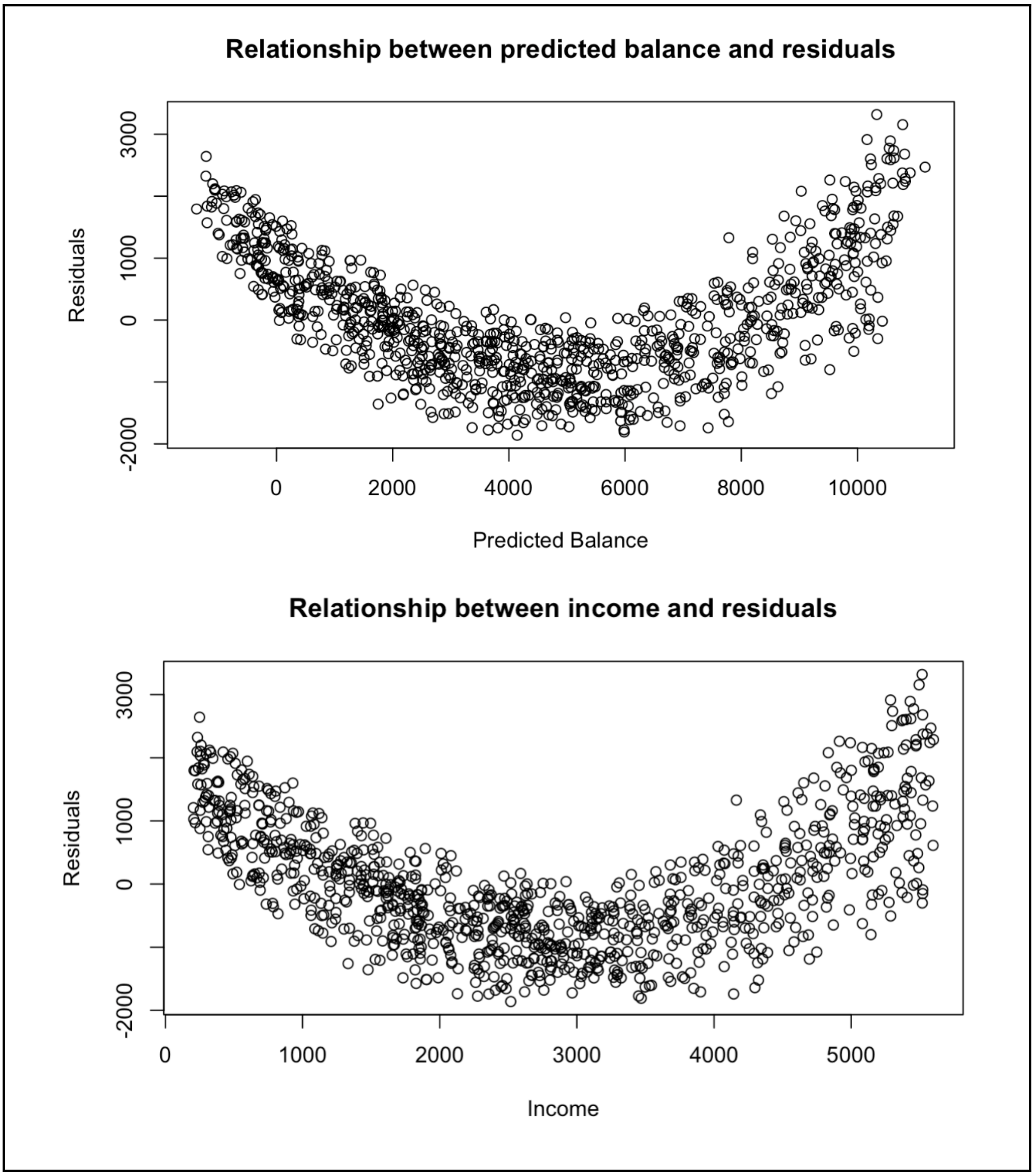


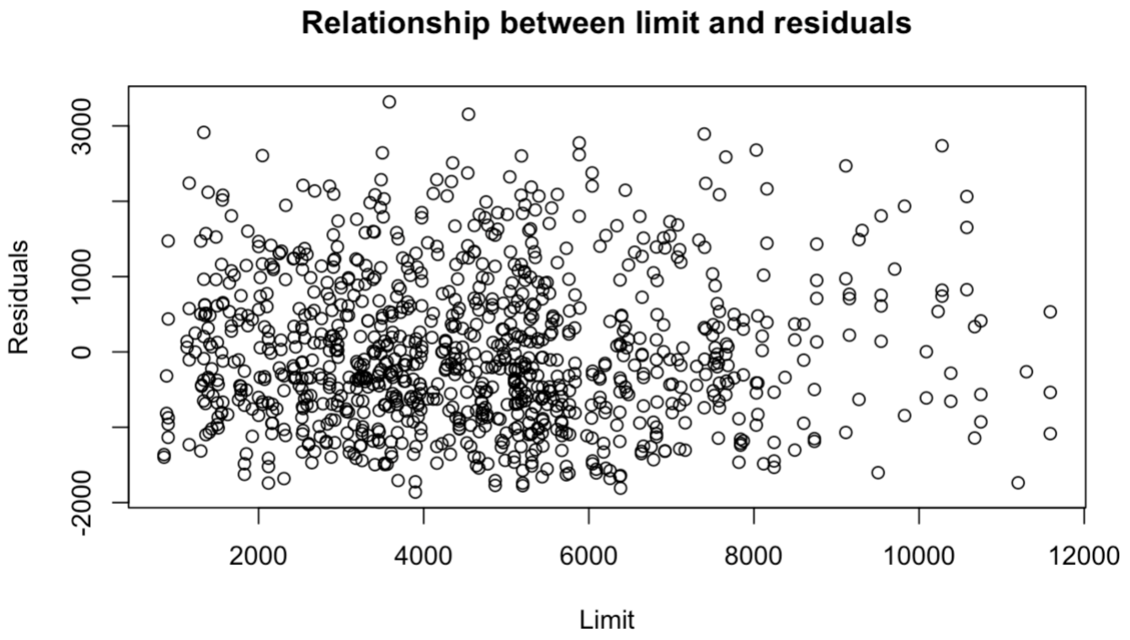
```
     Income          Limit            Rating            Cards             Age         Education          Gender     Married
 Min.   : 204   Min.   :  855   Min.   :103.0   Min.   :1.000   Min.   :44   Min.   : 6.00    Male :514   No :390
 1st Qu.:1447   1st Qu.: 3000   1st Qu.:232.8   1st Qu.:1.000   1st Qu.:54   1st Qu.:12.75   Female:486   Yes:610
 Median :2608   Median : 4438   Median :338.5   Median :2.000   Median :57   Median :14.00
 Mean   :2732   Mean   : 4605   Mean   :348.0   Mean   :2.456   Mean   :57   Mean   :13.88
 3rd Qu.:4042   3rd Qu.: 5758   3rd Qu.:431.5   3rd Qu.:3.000   3rd Qu.:60   3rd Qu.:15.00
 Max.   :5600   Max.   :11589   Max.   :825.0   Max.   :9.000   Max.   :69   Max.   :20.00
            Ethnicity       Balance
 African American:248   Min.   :   43.9
 Asian           :290   1st Qu.: 1650.9
 Caucasian       :462   Median : 3429.2
                        Mean   : 4497.6
                        3rd Qu.: 6718.2
                        Max.   :13933.7
[1] 16727.15
```

3.2.1 Input

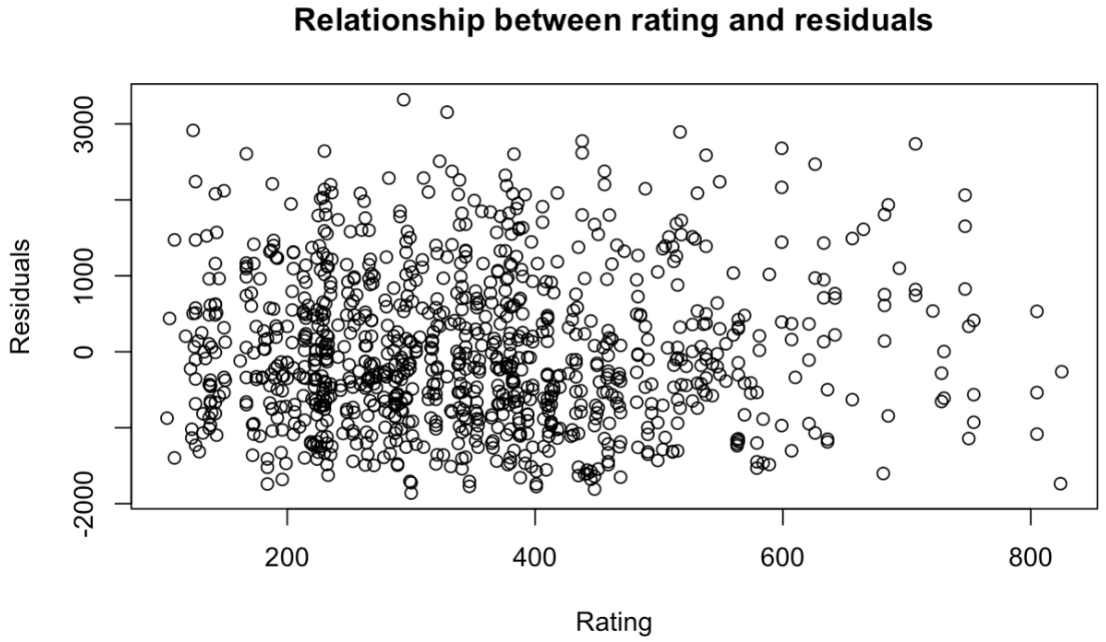
```
# Initial Model Assumptions
# Assumptions 1 and 4: Linearity, Homoskedastic errors

plot(initialModel$fitted.values, residuals(initialModel), main="Relationship between
predicted balance and residuals", xlab="Predicted Balance", ylab="Residuals")
plot(bank$Income, residuals(initialModel), main="Relationship between income
and residuals", xlab="Income", ylab="Residuals")

plot(bank$Limit, residuals(initialModel), main="Relationship between limit and
residuals",
xlab="Limit", ylab="Residuals")

plot(bank$Rating, residuals(initialModel), main="Relationship between rating and
residuals", xlab="Rating", ylab="Residuals")

plot(bank$Age, residuals(initialModel), main="Relationship between age and
residuals", xlab="Age", ylab="Residuals")

plot(bank$Education, residuals(initialModel), main="Relationship between
education and residuals", xlab="Education", ylab="Residuals")
```

3.2.1 Output

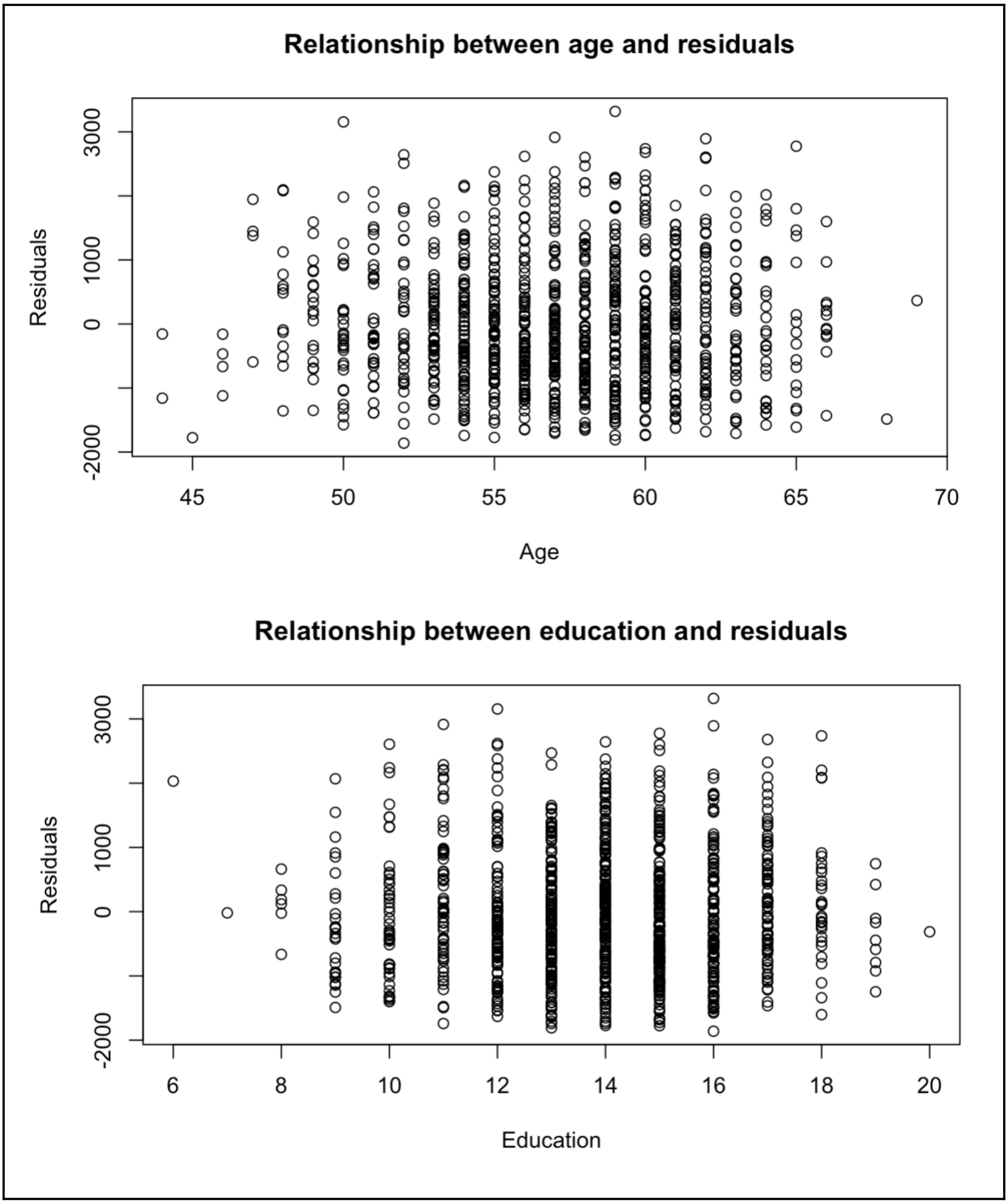

Relationship between predicted balance and residuals
Residuals
3000
1000
0
-2000
0
2000
4000
6000
8000
10000
Predicted Balance
Relationship between income and residuals
Residuals
3000
1000
0
-2000
0
1000
2000
3000
4000
5000
Income



Relationship between limit and residuals



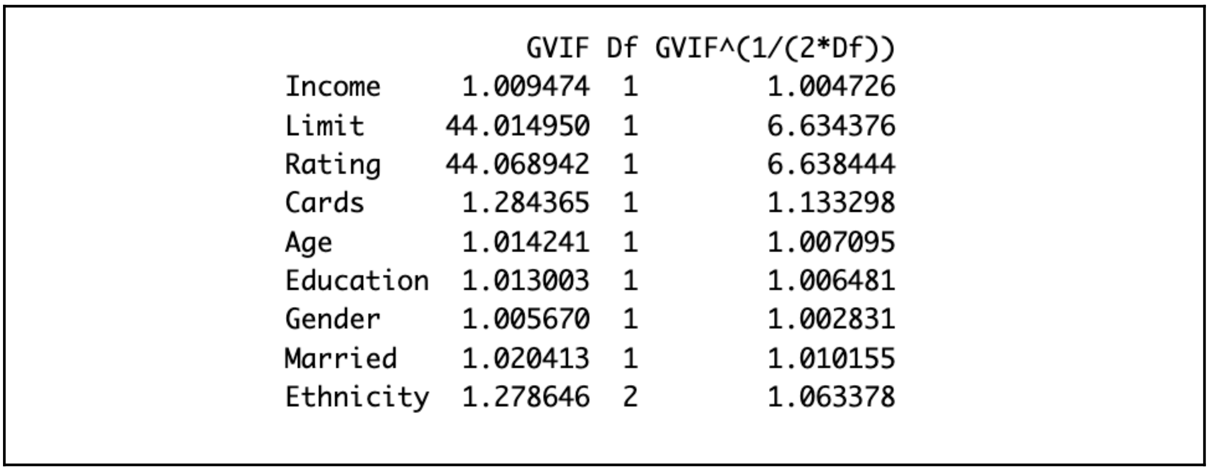
Relationship between rating and residuals



3.2.2 Input

```
# Assumption 2: Multicollinearity
library(car)
vif(initialModel)
```

3.2.2 Output

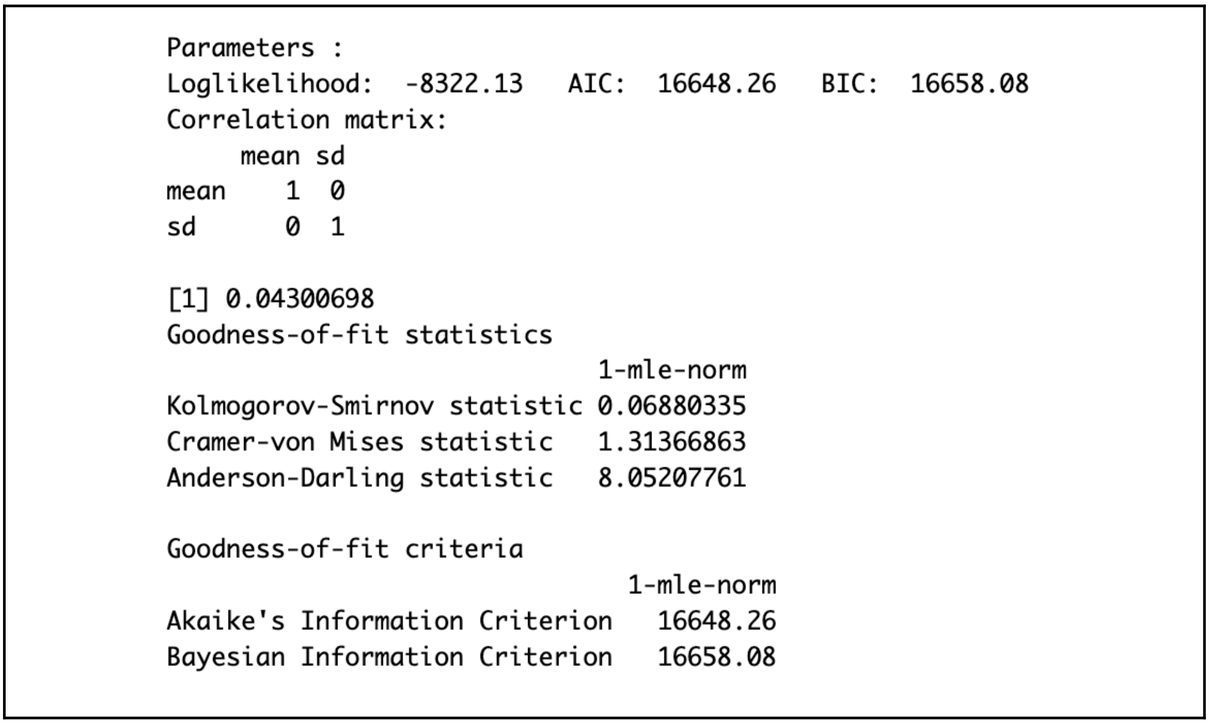


```
                 GVIF Df GVIF^(1/(2*Df))
Income       1.009474  1        1.004726
Limit       44.014950  1        6.634376
Rating      44.068942  1        6.638444
Cards        1.284365  1        1.133298
Age          1.014241  1        1.007095
Education    1.013003  1        1.006481
Gender       1.005670  1        1.002831
Married      1.020413  1        1.010155
Ethnicity    1.278646  2        1.063378
```

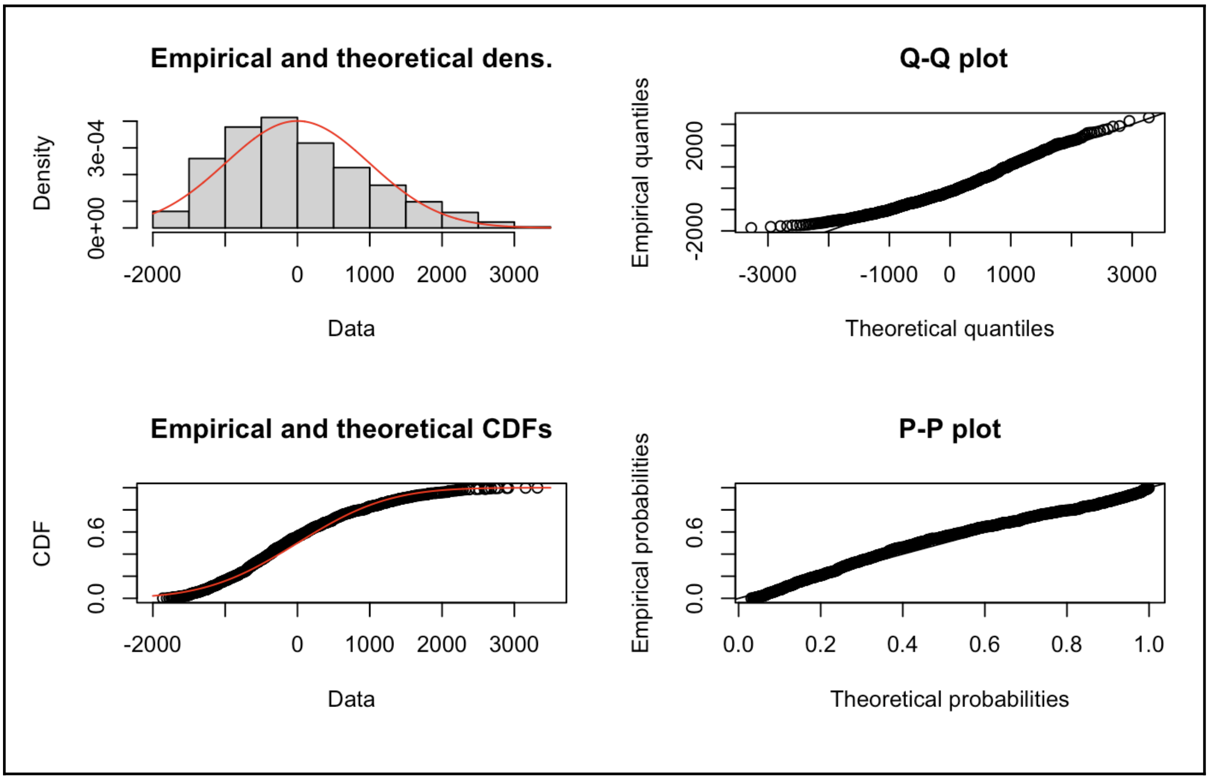
3.2.3 Input

```
# Assumption 5: Normally distributed errors
library(fitdistrplus)
fnorm_initialModel <- fitdist(residuals(initialModel), distr="norm")
summary(fnorm_initialModel)
plot(fnorm_initialModel)
KSCritValue <- 1.36/sqrt(1000) # for alpha = 0.05
KSCritValue
result1 <- gofstat(fnorm_initialModel, discrete=FALSE)
result1
```

3.2.3 Output



```
Parameters :
Loglikelihood:  -8322.13   AIC:  16648.26   BIC:  16658.08
Correlation matrix:
     mean sd
mean    1  0
sd      0  1

[1] 0.04300698
Goodness-of-fit statistics
                             1-mle-norm
Kolmogorov-Smirnov statistic 0.06880335
Cramer-von Mises statistic   1.31366863
Anderson-Darling statistic   8.05207761

Goodness-of-fit criteria
                               1-mle-norm
Akaike's Information Criterion   16648.26
Bayesian Information Criterion   16658.08
```



4.0 Input

```
library(car)
#Regression Model against all independent variables
bank1 = lm(Balance~., data = bank)
summary(bank1)

#Regression Model without age, married and ethnicity
bank2 = lm(Balance~.-Age-Married-Ethnicity, data = bank)
summary(bank2)

BIC(bank1)
BIC(bank2)
```

4.0 Output

```
Call:
lm(formula = Balance ~ ., data = bank)

Residuals:
    Min      1Q  Median      3Q     Max 
-1860.9  -713.5  -157.8   610.8  3319.0 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -580.24541  481.66366  -1.205  0.22862    
Income                2.14670    0.02047 104.861  < 2e-16 ***
```

```
Limit              -0.04219    0.09870  -0.427  0.66916
Rating              1.52757    1.50491   1.015  0.31033
Cards             -82.14049   26.70780  -3.076  0.00216 **
Age                -4.48177    7.62160  -0.588  0.55664
Education         -29.37544   14.29148  -2.055  0.04010 *
GenderFemale     -483.58532   63.51085  -7.614 6.18e-14 ***
MarriedYes        -76.26186   65.55573  -1.163  0.24498
EthnicityAsian     -7.59926   94.63717  -0.080  0.93602
EthnicityCaucasian   55.24816   87.67519   0.630  0.52874
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1001 on 989 degrees of freedom
Multiple R-squared:  0.9194,  Adjusted R-squared:  0.9185
F-statistic:  1127 on 10 and 989 DF,  p-value: < 2.2e-16



Call:
lm(formula = Balance ~ . - Age - Married - Ethnicity, data = bank)

Residuals:
    Min      1Q  Median      3Q     Max
-1879.3  -717.1  -157.8   618.0  3243.7

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -870.76050  226.84596  -3.839 0.000132 ***
Income          2.14712    0.02045 105.018  < 2e-16 ***
Limit          -0.02984    0.09502  -0.314 0.753583
Rating          1.35551    1.46223   0.927 0.354142
Cards         -79.09008   25.71041  -3.076 0.002154 **
Education     -28.98334   14.25717  -2.033 0.042329 *
GenderFemale -482.05023   63.38167  -7.606 6.57e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1000 on 993 degrees of freedom
Multiple R-squared:  0.9191,  Adjusted R-squared:  0.9187
F-statistic:  1881 on 6 and 993 DF,  p-value: < 2.2e-16

[1] 16727.15
[1] 16702.1
```

4.1 Input

```
#Regression Model without limit, age, married and ethnicity variable
bank3 <- lm(Balance~.-Limit-Age-Married-Ethnicity, data=bank)
summary(bank3)
```

```
#Regression Model without rating, age, married and ethnicity variable
bank4= lm(Balance~.-Rating-Age-Married-Ethnicity, data = bank)
summary(bank4)

BIC(bank3)
BIC(bank4)
```

4.1 Output

```
Call:
lm(formula = Balance ~ . - Limit - Age - Married - Ethnicity, 
    data = bank)

Residuals:
    Min      1Q  Median      3Q     Max 
-1885.9  -714.9  -159.8   626.0  3247.4 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -852.54429  219.20316  -3.889 0.000107 ***
Income          2.14715    0.02044 105.068  < 2e-16 ***
Rating          0.90297    0.24700   3.656 0.000270 ***
Cards         -78.10680   25.50743  -3.062 0.002257 ** 
Education     -29.04075   14.24953  -2.038 0.041813 *  
GenderFemale -481.75497   63.34596  -7.605 6.58e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 999.8 on 994 degrees of freedom
Multiple R-squared:  0.9191,  Adjusted R-squared:  0.9187 
F-statistic:  2260 on 5 and 994 DF,  p-value: < 2.2e-16


Call:
lm(formula = Balance ~ . - Rating - Age - Married - Ethnicity, 
    data = bank)

Residuals:
    Min      1Q  Median      3Q     Max 
-1892.6  -724.8  -156.7   626.7  3247.6 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -811.39362  217.60328  -3.729 0.000203 ***
Income          2.14748    0.02044 105.063  < 2e-16 ***
Limit           0.05698    0.01606   3.549 0.000405 ***
Cards         -74.69681   25.26807  -2.956 0.003188 ** 
Education     -28.97424   14.25615  -2.032 0.042378 *  
```

```
GenderFemale -480.90434   63.36514  -7.589 7.38e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1000 on 994 degrees of freedom
Multiple R-squared:  0.9191,  Adjusted R-squared:  0.9187
F-statistic:  2258 on 5 and 994 DF,  p-value: < 2.2e-16


[1] 16695.29
[1] 16696.06
```

4.2 Input

```
Model4 <- lm(Balance~Income+I(Income^2)+Gender+Cards+Rating, data=df)
summary(Model4)
BIC(Model4)
```

4.2 Output

```
Call:
lm(formula = Balance ~ Income + I(Income^2) + Gender + Cards +
    Rating, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1913.30  -404.00    15.02   425.57  1699.91

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.752e+02  8.260e+01  10.595  < 2e-16 ***
Income        9.532e-02  5.092e-02   1.872   0.0615 .
I(Income^2)   3.572e-04  8.595e-06  41.557  < 2e-16 ***
GenderFemale -4.813e+02  3.834e+01 -12.552  < 2e-16 ***
Cards        -6.228e+01  1.545e+01  -4.030 6.01e-05 ***
Rating        6.376e-01  1.492e-01   4.273 2.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 605.5 on 994 degrees of freedom
Multiple R-squared:  0.9703,  Adjusted R-squared:  0.9702
F-statistic:  6503 on 5 and 994 DF,  p-value: < 2.2e-16


[1] 15692.43
```

4.3.1 Input

```
#Resetting dataframe
df <- read.csv("Bank2023P.csv",stringsAsFactors = TRUE)
attach(df)

#Assigning outcomes for readability and ease of coding and accountability
a <- Income
b <- I(Income^2)
c <- Cards
d <- Rating
e <- Gender

#We create our linear regression model with all 9 interaction terms of degree 3.
#(excluding a*b Income and Income^2)
lm.3 <- lm(Balance~a*b*c + a*b*d + a*b*e + a*c*d + a*c*e + a*d*e + b*c*d + b*c*e
+ c*d*e + b*d*e,data=df,na.action = "na.fail")
summary(lm.3)

#We create our linear regression model with all 9 interaction terms of degree 2
#(excluding a*b Income and Income^2)
lm.4 <- lm(Balance~a*c + a*d +a*e + b*c + b*d + b*e + c*d + c*e + d*e
,data=df,na.action = "na.fail")
summary(lm.4)

#Best Interaction term Model 5
newb <-
lm(Balance~I(Income^2)+Income+Gender+Cards+Rating+Income*Gender +
I(Income^2)*Gender + Cards*Gender,data=df)
summary(newb)
BIC(newb)
```

4.3.1 Output

```
Call:
lm(formula = Balance ~ a * b * c + a * b * d + a * b * e + a *
    c * d + a * c * e + a * d * e + b * c * d + b * c * e + c *
    d * e + b * d * e, data = df, na.action = "na.fail")

Residuals:
     Min       1Q   Median       3Q      Max
-1044.94  -362.03   -10.37   350.31  1379.35

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.091e+02  3.129e+02   2.586  0.00985 **
a            9.377e-02  3.733e-01   0.251  0.80173
b            3.424e-04  1.343e-04   2.550  0.01092 *
c           -1.890e+02  1.135e+02  -1.666  0.09611 .
d           -3.418e-01  8.586e-01  -0.398  0.69066
eFemale     -9.096e+01  3.099e+02  -0.293  0.76922
a:b          7.997e-09  1.499e-08   0.533  0.59381
```

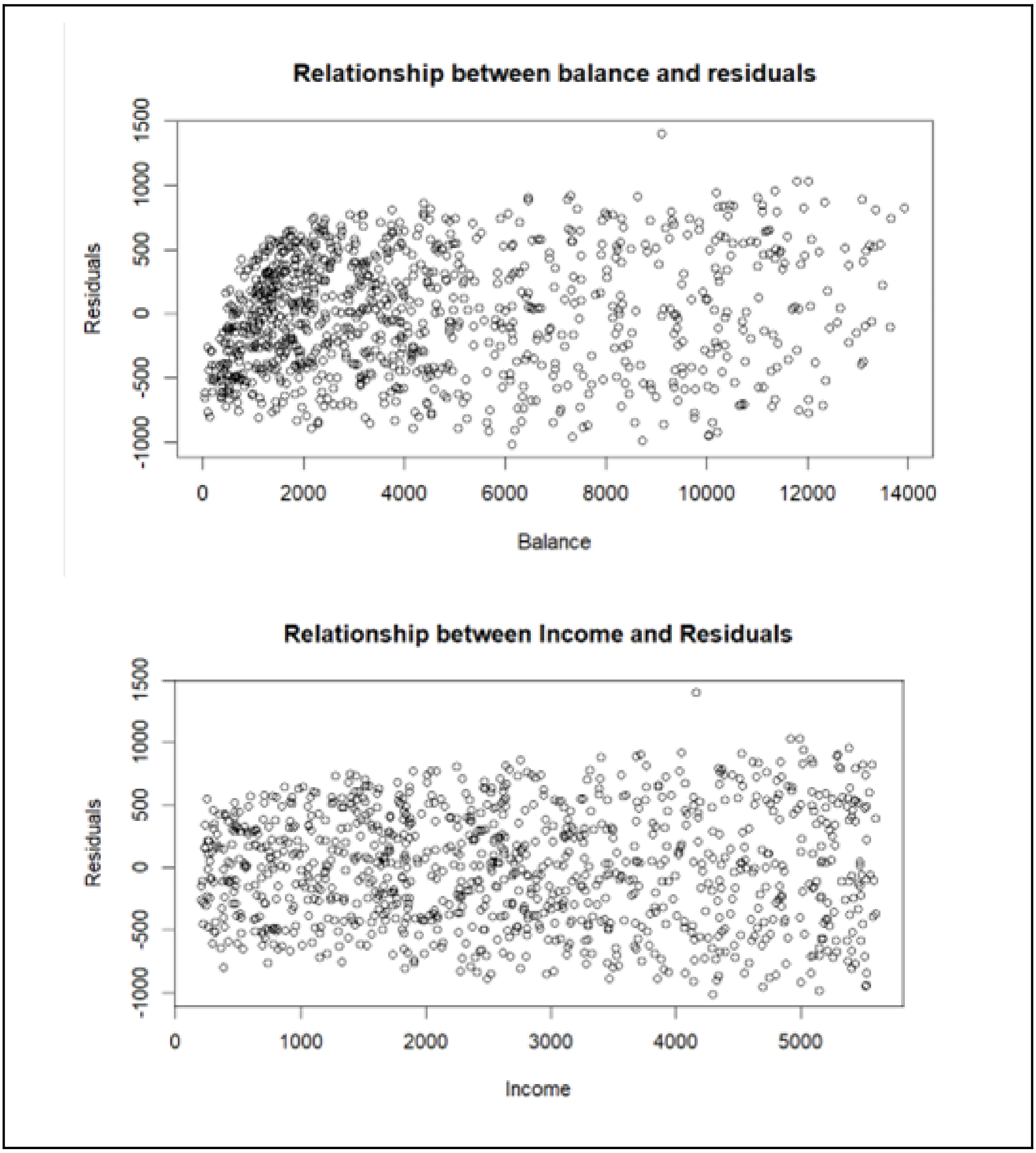
```
a:c          8.585e-02  1.186e-01   0.724  0.46931
b:c         -1.375e-05  3.749e-05  -0.367  0.71383
a:d          2.762e-04  9.977e-04   0.277  0.78195
b:d          1.081e-07  3.452e-07   0.313  0.75423
a:eFemale    3.089e-01  2.868e-01   1.077  0.28182
b:eFemale   -2.053e-04  8.915e-05  -2.303  0.02149 *
c:d          2.906e-01  2.314e-01   1.256  0.20944
c:eFemale    7.600e+01  1.001e+02   0.759  0.44793
d:eFemale    4.532e-01  7.996e-01   0.567  0.57100
a:b:c       -3.213e-10  3.902e-09  -0.082  0.93440
a:b:d       -2.380e-11  3.748e-11  -0.635  0.52567
a:b:eFemale  1.549e-08  9.677e-09   1.600  0.10985
a:c:d       -3.117e-04  1.943e-04  -1.605  0.10891
a:c:eFemale  2.795e-02  6.376e-02   0.438  0.66116
a:d:eFemale -2.727e-04  5.952e-04  -0.458  0.64700
b:c:d        5.944e-08  3.313e-08   1.794  0.07307 .
b:c:eFemale -6.499e-06  1.050e-05  -0.619  0.53606
c:d:eFemale -1.596e-02  1.417e-01  -0.113  0.91033
b:d:eFemale  1.921e-08  1.014e-07   0.189  0.84986
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 453.4 on 974 degrees of freedom
Multiple R-squared:  0.9837,   Adjusted R-squared:  0.9833
F-statistic:  2352 on 25 and 974 DF,  p-value: < 2.2e-16


Call:
lm(formula = Balance ~ a * c + a * d + a * e + b * c + b * d +
    b * e + c * d + c * e + d * e, data = df, na.action = "na.fail")

Residuals:
     Min       1Q   Median       3Q      Max
-1049.88  -361.92    -3.07   350.65  1394.93

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.115e+02  1.542e+02   3.966 7.85e-05 ***
a            1.919e-01  1.154e-01   1.663 0.096686 .
c           -9.552e+01  4.875e+01  -1.960 0.050327 .
d            2.596e-01  4.107e-01   0.632 0.527404
eFemale      2.349e+02  1.213e+02   1.937 0.053041 .
b            3.685e-04  1.981e-05  18.605  < 2e-16 ***
a:c         -1.154e-02  3.087e-02  -0.374 0.708540
a:d         -8.639e-05  2.924e-04  -0.295 0.767703
a:eFemale   -2.338e-02  7.645e-02  -0.306 0.759814
c:b          2.757e-06  5.101e-06   0.540 0.589003
d:b          4.808e-08  4.967e-08   0.968 0.333192
eFemale:b   -8.246e-05  1.290e-05  -6.394 2.49e-10 ***
c:d          2.014e-02  6.894e-02   0.292 0.770262
c:eFemale    7.970e+01  2.369e+01   3.364 0.000797 ***
d:eFemale   -8.829e-02  2.267e-01  -0.389 0.697071
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 453.5 on 985 degrees of freedom
Multiple R-squared:  0.9835,   Adjusted R-squared:  0.9833
F-statistic:  4196 on 14 and 985 DF,  p-value: < 2.2e-16
```

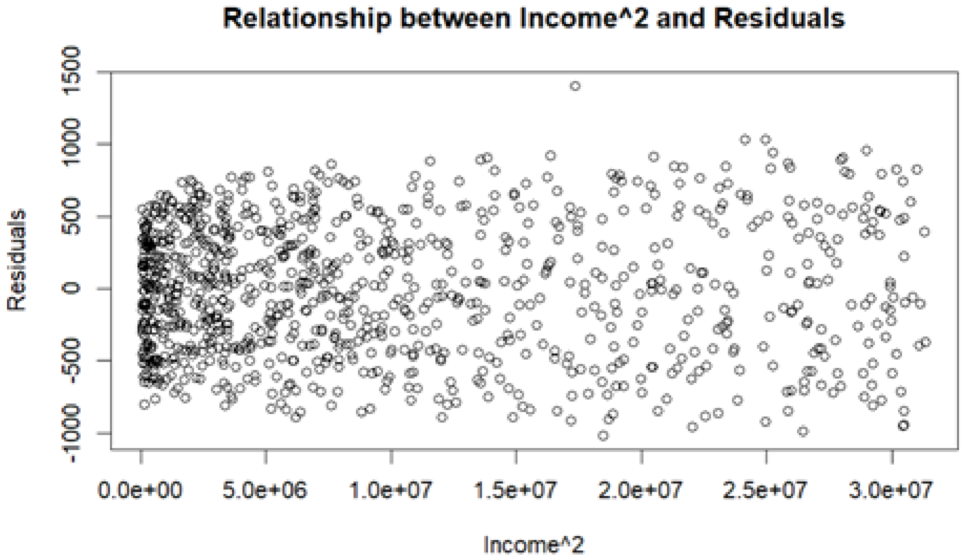
```
Call:
lm(formula = Balance ~ I(Income^2) + Income + Gender + Cards + 
    Rating + Income * Gender + I(Income^2) * Gender + Cards * 
    Gender, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1016.5  -376.1    10.4   364.1  1403.4 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               5.540e+02  7.954e+01   6.964 6.01e-12 ***
I(Income^2)               3.945e-04  9.064e-06  43.519  < 2e-16 ***
Income                    1.214e-01  5.395e-02   2.250 0.024655 *  
GenderFemale              1.960e+02  1.083e+02   1.810 0.070658 .  
Cards                    -9.325e+01  1.493e+01  -6.248 6.18e-10 ***
Rating                    5.030e-01  1.124e-01   4.475 8.51e-06 ***
Income:GenderFemale      -1.773e-02  7.656e-02  -0.232 0.816945    
I(Income^2):GenderFemale -8.331e-05  1.291e-05  -6.451 1.74e-10 ***
GenderFemale:Cards        7.914e+01  2.160e+01   3.665 0.000261 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 454.9 on 991 degrees of freedom
Multiple R-squared:  0.9833,	Adjusted R-squared:  0.9832 
F-statistic:  7298 on 8 and 991 DF,  p-value: < 2.2e-16

[1] 15138.18
```

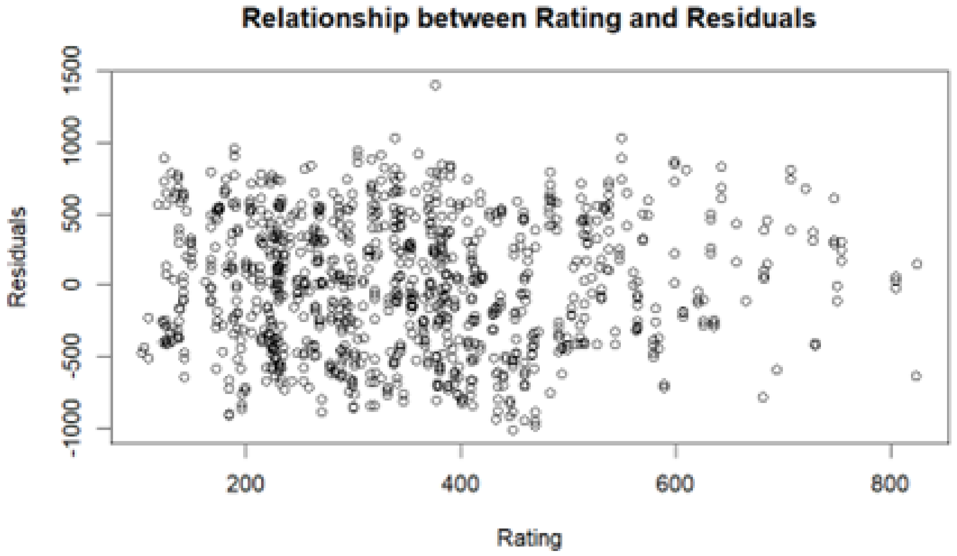
5.1 Input

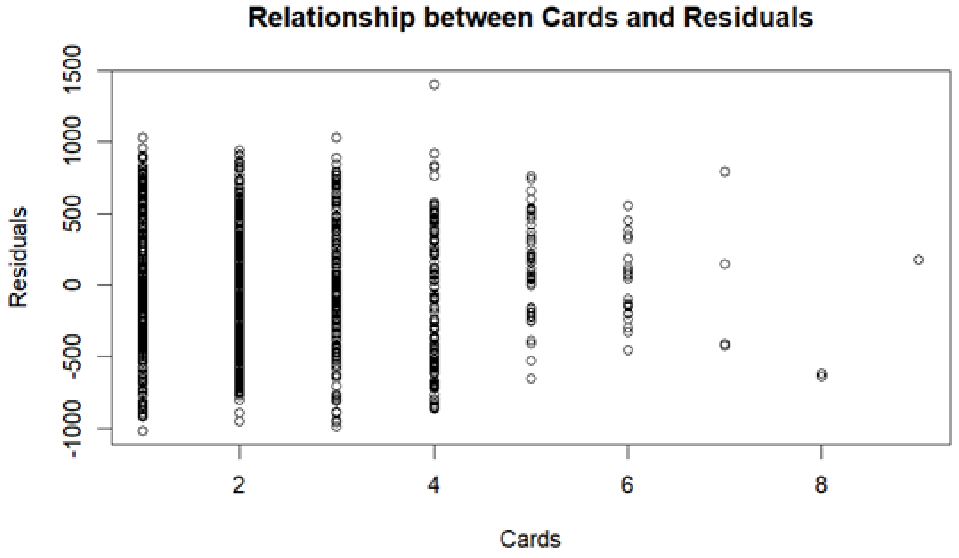
```
# Proposed Model Assumptions
# Assumptions 1 and 4: Linearity, Homoskedastic errors
lm.best <-
lm(Balance~I(Income^2)+Income+Gender+Cards+Rating+Income*Gender +
I(Income^2)*Gender + Cards*Gender,data=df)
bestsid <- residuals(lm.best)
plot(Balance, bestsid, main = "Relationship between balance and residuals",
    xlab="Balance", ylab = "Residuals")

plot(Income, bestsid, main="Relationship between Income and Residuals",
xlab="Income", ylab="Residuals")
plot(Income^2, bestsid, main="Relationship between Income^2 and
Residuals",xlab="Income^2", ylab="Residuals")
plot(Rating, bestsid, main="Relationship between Rating and Residuals",
xlab="Rating", ylab="Residuals")
plot(Cards, bestsid, main="Relationship between Cards and Residuals",
xlab="Cards", ylab="Residuals")
```

5.1 Output


Relationship between balance and residuals
Residuals
-1000
-500
0
500
1000
1500
0
2000
4000
6000
8000
10000
12000
14000
Balance
Relationship between Income and Residuals
Residuals
-1000
-500
0
500
1000
1500
0
1000
2000
3000
4000
5000
Income


Relationship between Income^2 and Residuals
Residuals
1500
1000
500
0
-500
-1000
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Income^2

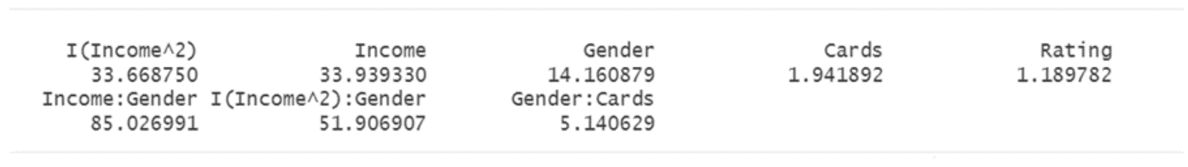

Relationship between Rating and Residuals
Residuals
1500
1000
500
0
-500
-1000
200
400
600
800
Rating



5.2 Input

```
# Assumption 2: Multicollinearity
vif(lm.best)
```

5.2 Output



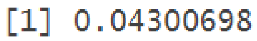
```
  I(Income^2)            Income           Gender            Cards           Rating 
    33.668750         33.939330        14.160879         1.941892         1.189782 
Income:Gender I(Income^2):Gender     Gender:Cards 
    85.026991         51.906907         5.140629 
```

5.3 Input

```
#Assumption 5: Normally distributed errors
library(fitdistrplus)
KScritvalue <-1.36/sqrt(1000)    #at 5%
KScritvalue
fnorm <- fitdist(bestsid, "norm")
res1 <- gofstat(fnorm, discrete = FALSE)
res1
plot(fnorm)
```

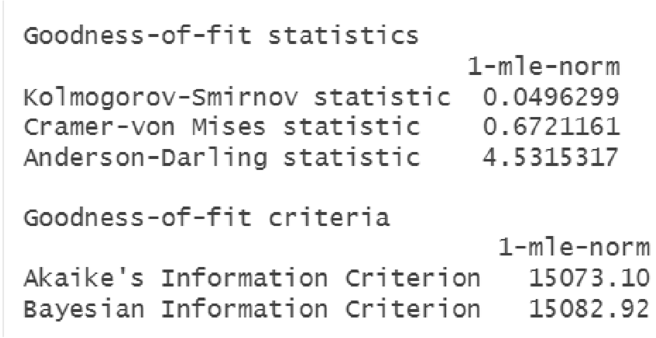
5.3 Output

KS critical value:



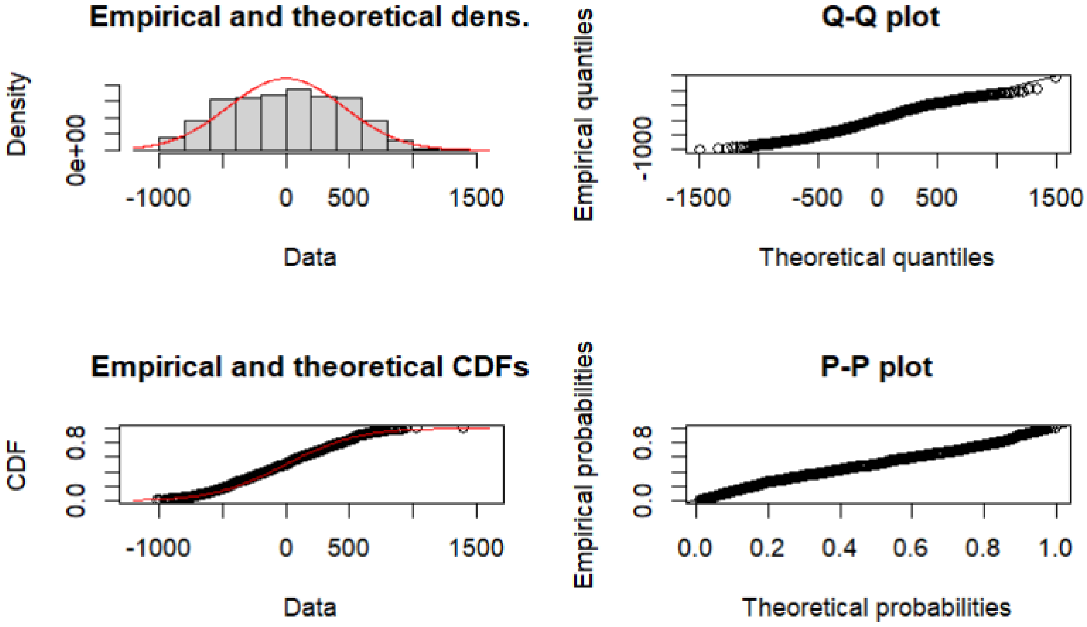
```
[1] 0.04300698
```

KS statistic:



```
Goodness-of-fit statistics
                             1-mle-norm
Kolmogorov-Smirnov statistic  0.0496299
Cramer-von Mises statistic    0.6721161
Anderson-Darling statistic    4.5315317

Goodness-of-fit criteria
                               1-mle-norm
Akaike's Information Criterion   15073.10
Bayesian Information Criterion   15082.92
```

P-P and Q-Q plot:


Empirical and theoretical dens.
Density
0e+00
Data
Q-Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P-P plot
Empirical probabilities
Theoretical probabilities