****

**DSA211-Statistical Learning with R-G1**

**Group Project Part 1\_Group 6**

**Instructor: Goh Jing Rong**

**Group Members:**

|  |
| --- |
| \*Censored\* |
| \*Censored\* |
| Chong Gao Zhe |
| \*Censored\* |
| \*Censored\* |

Contents

[**1. Introduction** 3](#_Toc116832619)

[**1.1 Objective - Constructing a multi-variable regression model to predict Balance** 3](#_Toc116832620)

[**1.2 Hypothesis** 3](#_Toc116832621)

[**2. Exploratory Data Analysis** 3](#_Toc116832622)

[**2.1 Exploratory Data Analysis Overview** 3](#_Toc116832623)

[**2.2 Relationships between all Variables** 4](#_Toc116832624)

[**2.2.1 Scatter Plots for Quantitative Variables** 4](#_Toc116832625)

[**2.2.2. Boxplots for Categorical Variables** 4](#_Toc116832626)

[**3. Model Selection** 5](#_Toc116832627)

[**3.1 Selection Criterion** 5](#_Toc116832628)

[**3.2 Multicollinearity Problem** 5](#_Toc116832629)

[**3.3 Linear Model Exploration** 5](#_Toc116832630)

[**3.3.1 Model 1 (Balance ~ .)** 5](#_Toc116832631)

[**3.3.2 Model 2 (Balance ~ Income + Gender)** 5](#_Toc116832632)

[**3.4 Polynomial Model Exploration** 6](#_Toc116832633)

[**3.4.1 Model 3 (Balance~ poly(Income,2,raw = TRUE) + Gender)** 6](#_Toc116832634)

[**3.4.2 Higher-order polynomials** 6](#_Toc116832635)

[**3.4.3 Model 4 (Balance~ poly(Income,2,raw = TRUE) + Gender + Cards + Rating)** 6](#_Toc116832636)

[**3.5 Model with Interaction Terms Exploration** 7](#_Toc116832637)

[**3.5.1 Model 5: Investigating Interaction Terms of Model 4** 7](#_Toc116832638)

[**3.5.2 Model 6 (Balance~poly(Income, 2,raw=TRUE)\*Gender + Cards\*Gender + Rating\*Gender** 7](#_Toc116832639)

[**3.6 Final Model Exploration** 7](#_Toc116832640)

[**3.6.1 Model 7 (Balance~poly(Income, 2,raw=TRUE)\*Gender + Cards\*Gender + Rating)** 7](#_Toc116832641)

[**4. Best Model (Model 7) Assumption Justifications** 8](#_Toc116832642)

[**5. Joint Hypothesis Testing for Best Model (Model 7)** 8](#_Toc116832643)

[**6. Appendix** 9](#_Toc116832644)

# **1. Introduction**

## **1.1 Objective - Constructing a multi-variable regression model to predict Balance**

The objective of the project is to construct a multi-variable regression model to predict the bank account balance (*Balance*) of an individual using independent variables such as: monthly salary (*Income*), credit limit (*Limit*), credit rating score (*Rating*), number of credit cards (*Cards*), age (*Age*), years of education (*Education*), gender (*Gender*), marital status (*Married*), and ethnic group (*Ethnicity*).

## **1.2 Hypothesis**

The team’s initial hypothesis is that *Income* will have the most impact on the model as an independent variable. As an individual’s monthly salary increase, we expect that the individual will have a higher disposable income. Therefore, assuming that the savings as a proportion of an individual’s income remain relatively the same or increase when disposable income increases, there will be a higher absolute amount of savings in an individual's bank account.

Another hypothesis that we had was on the potential interaction effect between the independent variables. Of the independent variables, we suspected both *Gender* and *Education* could have a potential effect on *Income*. According to a study by the Ministry of Manpower (MOM) and National University of Singapore (NUS), the adjusted gender pay gap was 6% in 2018, meaning that a female is likely to earn 6% less than a male when all else is equal. Also, research conducted by Organisation for Economic Co-operation and Development (OECD) in 2016 found that Singapore has the highest percentage change in wages as years of education increased when compared to 33 other economies. Therefore, we look to explore these effects in our models below.

# **2. Exploratory Data Analysis**

## **2.1 Exploratory Data Analysis Overview**

After forming our hypotheses, we set out to explore the data provided to understand them better and plausibly gain additional insights. We decided to do this through finding the relationship between the quantitative as well as qualitative variables against *Balance*.

## **2.2 Relationships between all Variables**

To assist in our decision on the best independent variables to include in our multiple regression model, we first conduct an initial visualisation of the relationships between the dependent variable Balance and the nine independent variables, using scatterplots for quantitative variables such as Income, Limit, Rating and Age, while using boxplots for categorical variables such as Cards, Education, Gender, Married and Ethnicity.

## **2.2.1 Scatter Plots for Quantitative Variables**

Using the pairs() function for Balance and the four quantitative variables, we obtain the following scatter plots. From Figure 1, Balance does not seem to have any apparent relationship with Limit, Rating and Age. On the contrary, Income seems to have a quadratic relationship with Balance, and hence we would consider exploring Income polynomials of higher orders as well. In addition, the scatter plot of Limit and Rating hints at a linear relationship. In fact, looking at Figure 2, we see that the correlation between Limit and Rating is 0.98895794, which is close to 1. This means that Limit and Rating are highly correlated and should not be included together in our regression model, as it could result in imperfect multicollinearity. We can also notice that Income greatly affects Balance, and would be an important variable to include in our model.

## **2.2.2. Boxplots for Categorical Variables**

After plotting the boxplots for the categorical variables in Figure 3, we found that the median and boxplots of *Gender*, *Ethnicity*, and *Marital Status* to be relatively similar at different values / levels. Hence, this could mean that the 3 categorical variables above are likely to be less important, contrary to our hypothesis on *Gender*. On the other hand, we found that the median *Balance* tends to vary at different levels of *Education* and *Cards*, which means that *Education* and *Cards* could potentially be more important variables than the other categorical variables.

# **3. Model Selection**

## **3.1 Selection Criterion**

Looking at adjusted R-squared may be a good way of quantifying the extent to which regressors account for, or explain, the variation in the dependent variable. Nevertheless, heavy reliance on the adjusted R squared or R squared can be a trap. Hence, the BIC criterion is better for comparing between models as it places greater penalty on models with more variables, allowing us to select the more parsimonious model and avoid overfitting. Additionally, we would attempt to drop predictors with insignificant p-values, provided it decreases the BIC.

## **3.2 Multicollinearity Problem**

When two or more regressors are highly correlated in the sense that there is a linear function of the regressors that is highly correlated with another regressor, it leads to imperfect multicollinearity, resulting in the coefficients of the regressors to be imprecisely estimated; leading to a large sampling variance.

## **3.3 Linear Model Exploration**

### **3.3.1 Model 1 (Balance ~ .)**

For the first model, we regress Balance on all the independent variables. We observed Income and Gender to be statistically significant at 1% level of significance while the other variables are not. The residual plot of residuals and Balance also shows that the homoskedasticity assumption is violated as there is a quadratic relationship involved.

**Model 1 BIC:** (16673.31)

### **3.3.2 Model 2 (Balance ~ Income + Gender)**

For our second model, we regress Balance on Income and Gender for further testing of their significance in explaining variation of Balance and use it as a baseline for adding additional regressors. As a result, Income and Gender are indeed statistically significant and violation of homoscedasticity assumption still remains.

**Model 2 BIC:** (16632.15)

## **3.4 Polynomial Model Exploration**

### **3.4.1 Model 3 (Balance~ poly(Income,2,raw = TRUE) + Gender)**

In our previous models, we observed a U-shaped relationship between Balance and Residuals suggesting that a quadratic regression would better fit the set of data as seen in Figure 4. Hence, we increase the polynomial order for Income to 2. As a result, the BIC did decrease from 16632.15 to 15718.36, suggesting that polynomial regression is better than linear in capturing the relationship.

**Model 3 BIC:** (15718.36)

### **3.4.2 Higher-order polynomials**

After recognizing the significance of polynomial regression in this set of data, we increase the order of polynomials and decide which is the best fit model for this data. As a result, the BIC did increase for third order and fourth order, suggesting that second order polynomial is better fit for the data,  
Therefore, we decided to proceed our further analysis with 2nd order polynomials.

**3rd order polynomial BIC:** (15723.3), **4th order polynomial BIC:** (15729.14)

### **3.4.3 Model 4 (Balance~ poly(Income,2,raw = TRUE) + Gender + Cards + Rating)**

Since we know that Income and Income^2 must be included into the regression model, we executed a best subset selection (Balance~I(Income^2) + .), and force.in Income and Income^2 into the regsubsets() function. Through this process, we obtain the best predictors according to lowest BIC which are Income, Income^2, Gender, Cards and Rating. Consequently, BIC has further decreased as compared to Model 2. Note that Education is still not included in the regression as BIC would increase to 15709.94, which could mean Education is not essential to our analysis as we had previously hypothesised.

**Model 4 BIC:** (15705.44)

## **3.5 Model with Interaction Terms Exploration**

### **3.5.1 Model 5: Investigating Interaction Terms of Model 4**

Now, we would like to regress and check on 5 predictors to see if it can better explain Balance. To ensure that we did not mistakenly exclude Education, we added Education in the multiple regression model to check the p-values of interaction terms. Through the regression model, we noticed that Income:Gender and Cards:Gender have extremely low p-values (5.20e-10 & 2.36e-07 respectively) and should be included into the model. In addition, Rating:Gender is slightly significant with a p-value of 0.0340. On the contrary, Education has a high p-value of 0.4783 and interaction terms involving Education are not significant at 5% level, which further supports our observation in Model 4 that the data does not provide sufficient evidence of Education being significantly related to Balance.

### **3.5.2 Model 6 (Balance~poly(Income, 2,raw=TRUE)\*Gender + Cards\*Gender + Rating\*Gender**

We do a multiple regression on those variables that were noted as significant in Model 5. As a result, we obtained a BIC of 15144.71 which were significantly lower than Model 4 and 5.

**Model 6 BIC:** (15144.71)

## **3.6 Final Model Exploration**

### **3.6.1 Model 7 (Balance~poly(Income, 2,raw=TRUE)\*Gender + Cards\*Gender + Rating)**

Observing that Rating:Gender only has a slight impact on Balance from its p-value as compared to the other interaction terms, we attempted to add only Rating in without its interaction term. This further decreases the BIC to 15140.98, implying that Rating:Gender are not significantly enough for us to include into the model.

**Model 7 BIC:** (15140.98)

# **4. Best Model (Model 7) Assumption Justifications**

Through our analysis, we have come to the conclusion that Model 7 is the best model according to the BIC criterion. To check for error constant variance assumption of the multiple regression model, we performed residual analyses of Model 7’s residuals against fitted values(Balance), as well as all independent variables involved. From the Figure 5, the constant variance assumption is not violated as their errors are fairly distributed across the values of our variables.

To check for normality assumption of our model, we use the Q-Q plot to assess if our residual describes a normal distribution and P-P plot in Figure 6 below for comparing our residuals to normal CDF less than or equal to each observed value against the actual proportion. As a result, we observe no violation of normality assumption. Moreover, the KS statistics is 0.05173395 < KS Critical value of 0.4300698. Hence, we do not have sufficient evidence to show the residual is not normally distributed.

# **5. Joint Hypothesis Testing for Best Model (Model 7)**

Model 7: Balance = 5.966e+02 + 1.224e-01 Income + 3.958e-04 Income^2 + 2.022e+02 Gender(Female) - 1.089e+02 Cards + 4.121e-01 Rating - 2.912e-02 Income:Gender(Female) - 8.328e-05 Income^2:Gender(Female) + 1.015e+02 Gender(Female):Cards

H0: All coefficients of independent variables = 0

H1: Not all coefficients of independent variables = 0

From the summary of model 7 in R, the p-value of overall F-test is < 2.2e-16 < 0.01 significance level, and we reject the null hypothesis. The data provides sufficient evidence of at least 1 predictor having a significant relationship with Balance at the 1% significance level.

# **6. Appendix**

**Hypotheses**

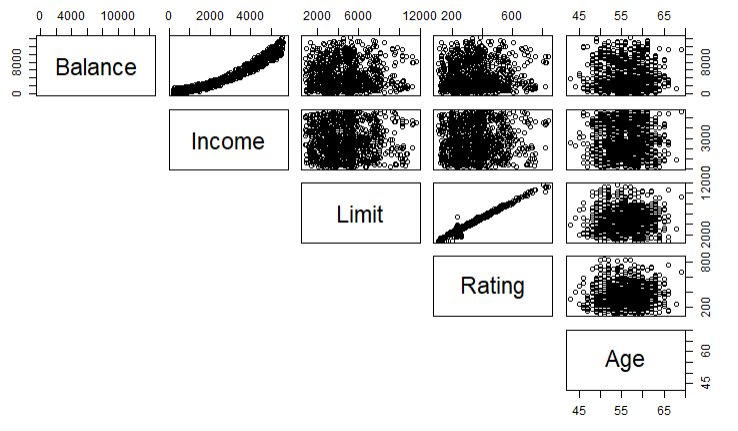
Adjusted Gender Pay Gap Research by NUS and MOM

Percentage Change of Wage as Education Increase by OECD

**Exploratory Data Analysis**

**Scatter Plots for Quantitative Variables**

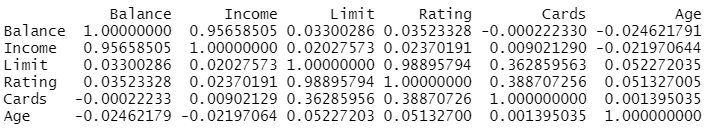
****



*Figure 1: Scatter Plots between Balance and Quantitative variables*

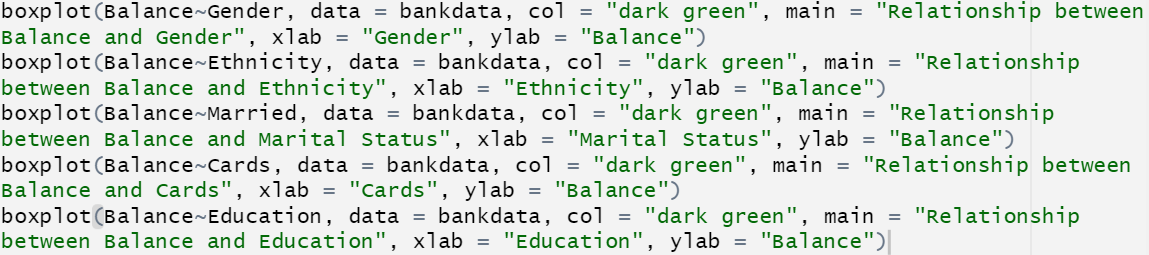
**Correlation for Balance, Income, Limit, Rating, Cards & Age**

****

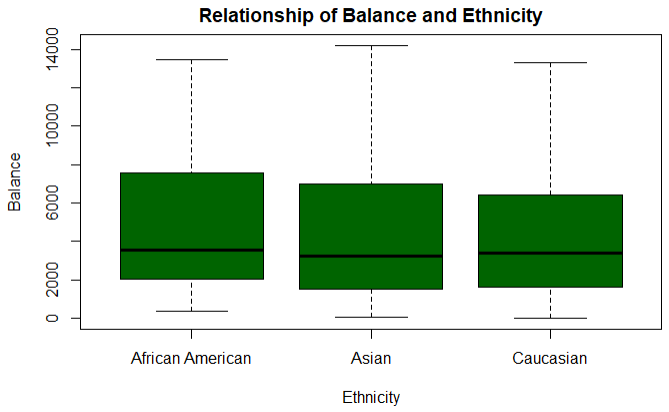


*Figure 2: Correlation of Balance, Income, Limit, Rating, Cards and Age*

**Boxplots for Categorical Variables**

****

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

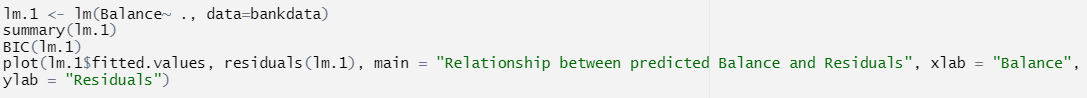
Description automatically generatedChart, box and whisker chart

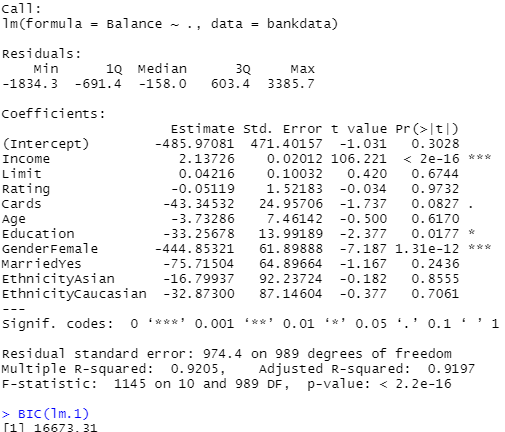
Description automatically generated

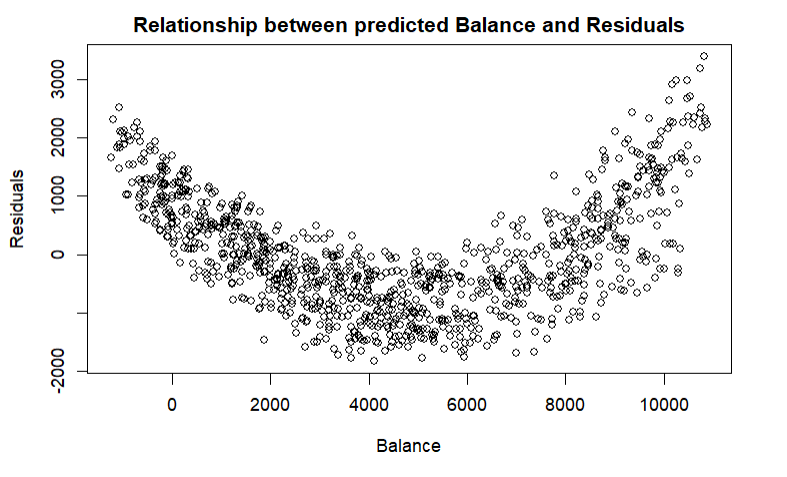
*Figure 3: Boxplot of Balance against Gender, Education, Ethnicity, Marital Status, Cards*

**Model Selection**

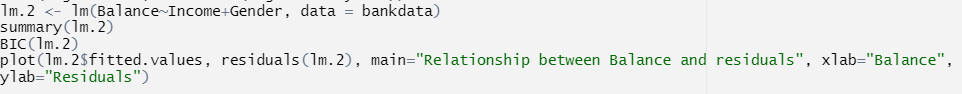
**Model 1**

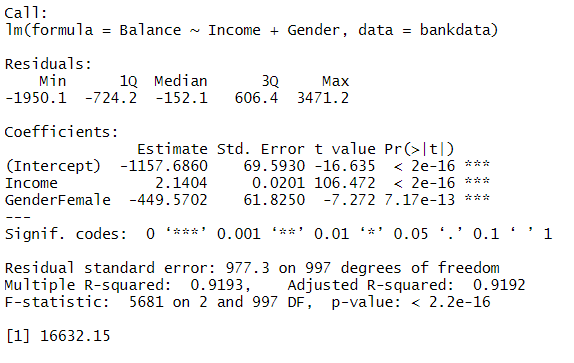
****

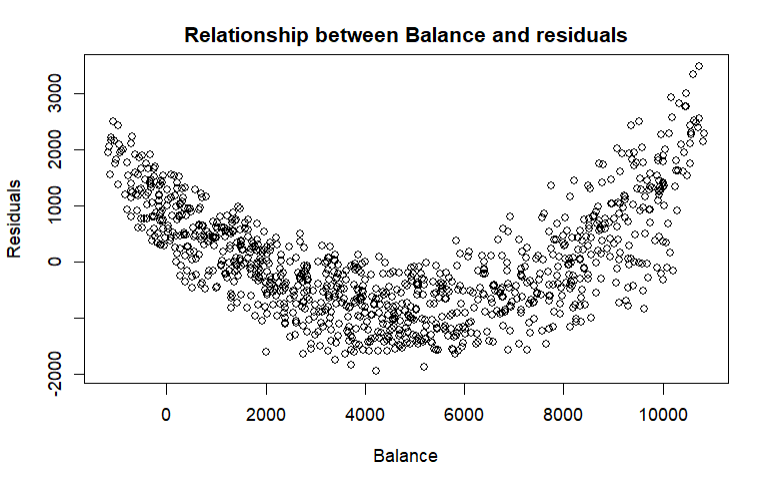
****

****

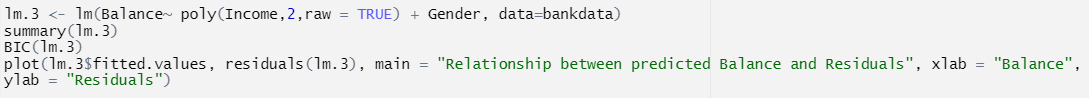
**Model 2**

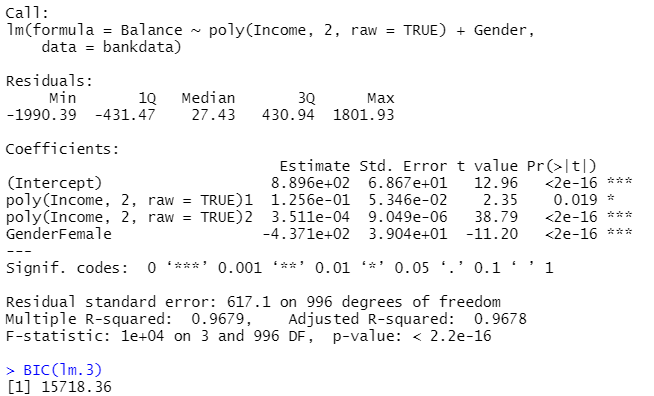
****

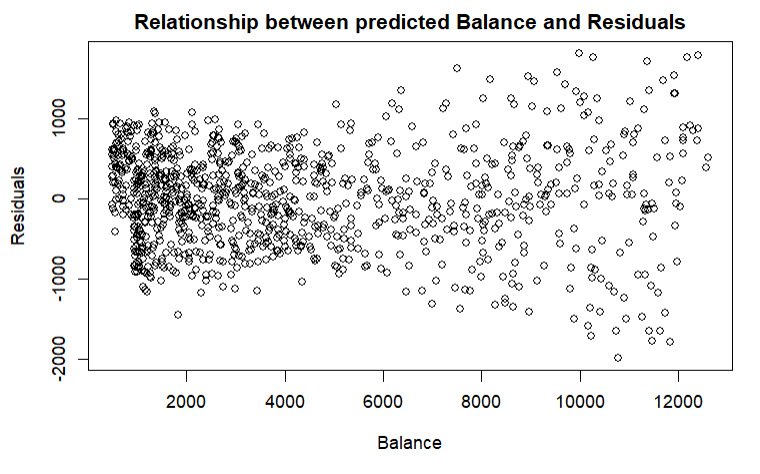
****

****

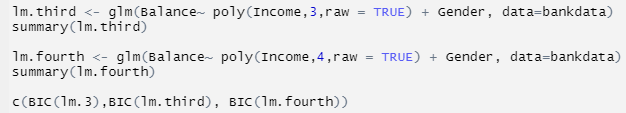
**Model 3**

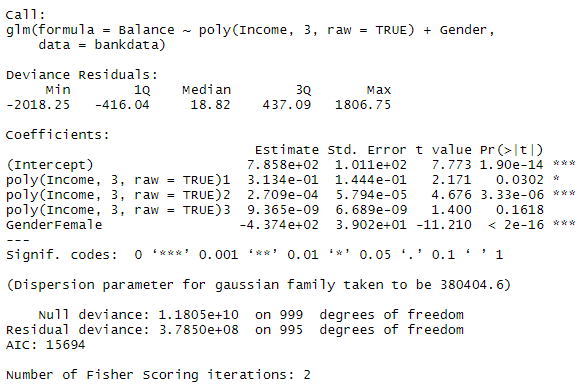
****

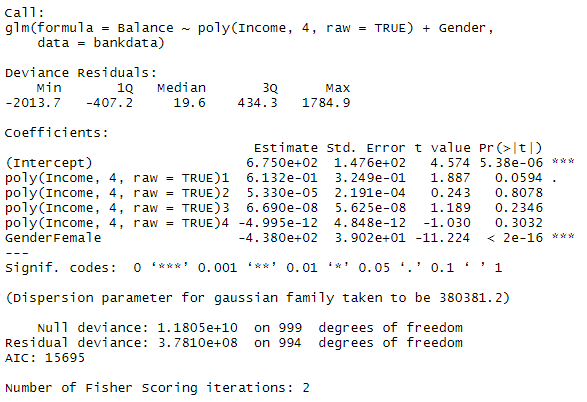
****

****

**Higher Order polynomials**

****

****

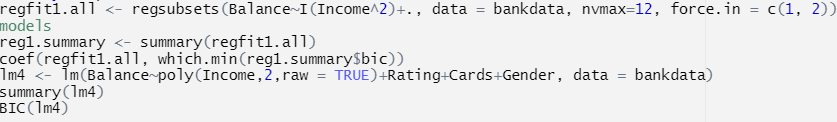
****

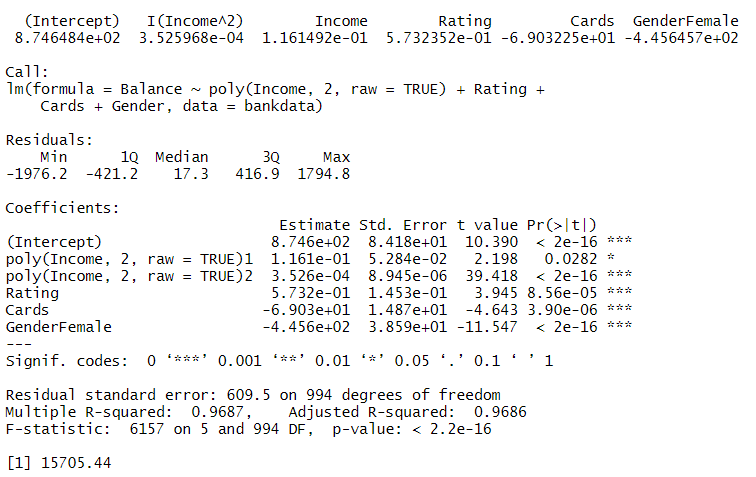
****



*Figure 4: Before and After Residual Plots*

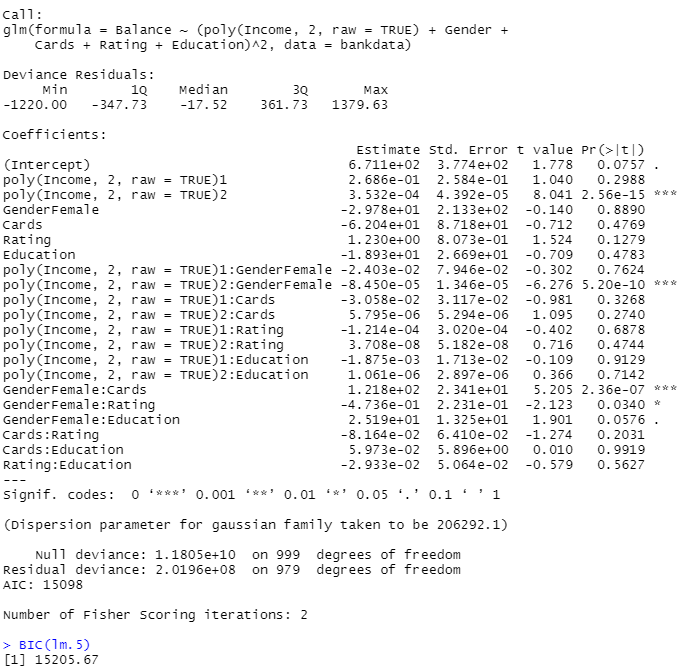
**Model 4**

****

****

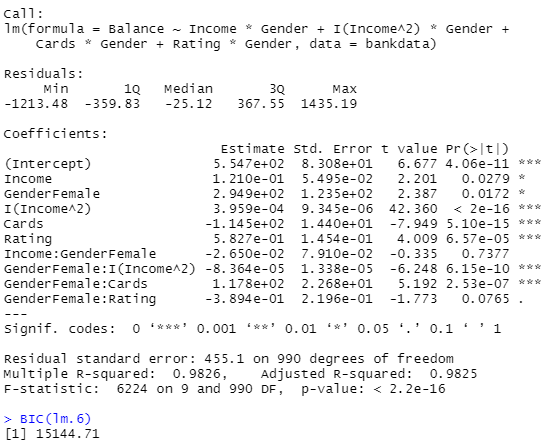
**Model 5**

****

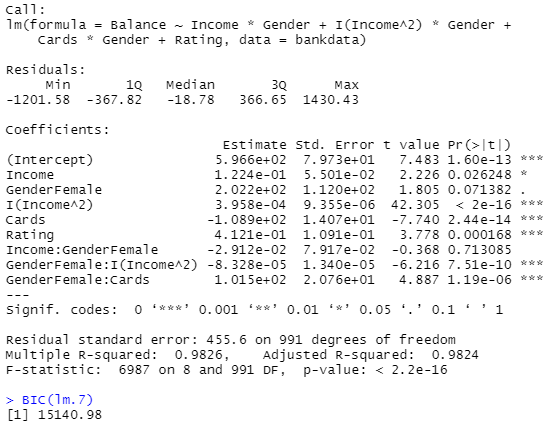
****

**Model 6:**

****

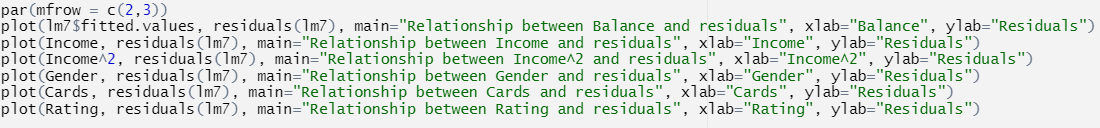
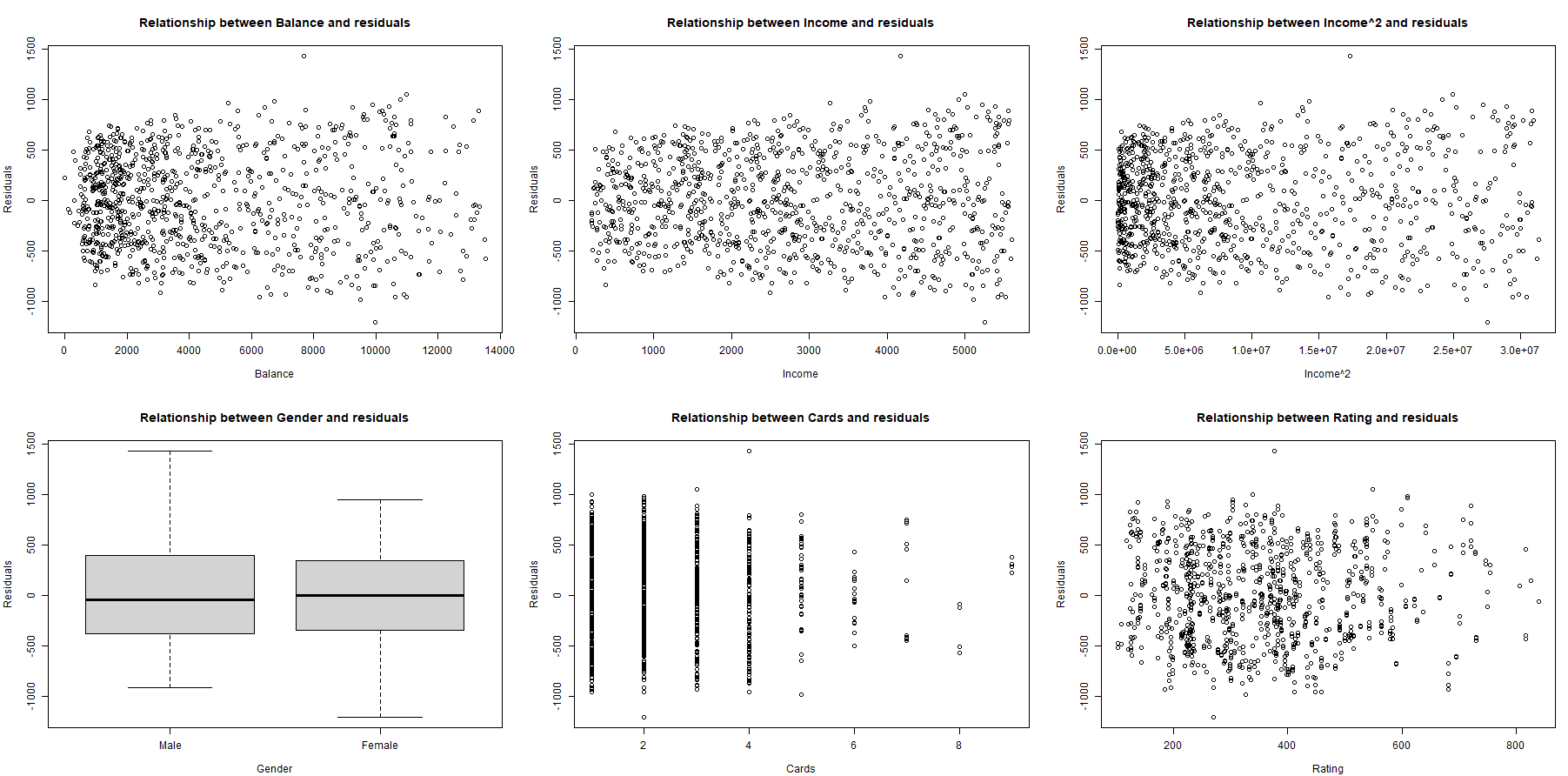
****

**Model 7:**

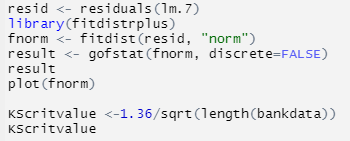
****

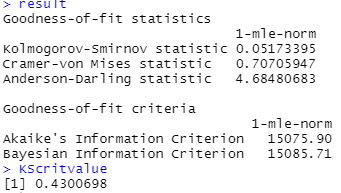
**Multiple Regression Assumption Justification**

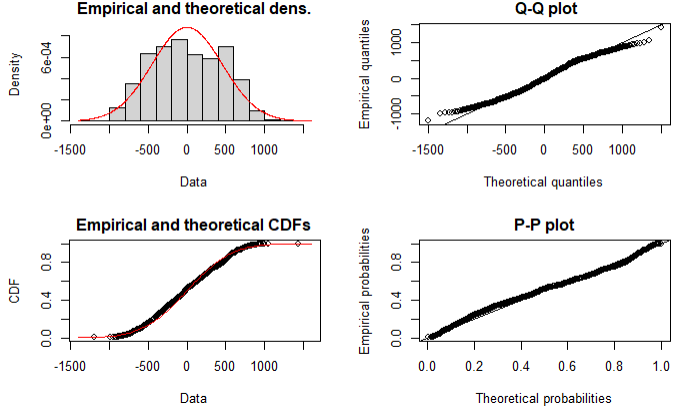
**Residual Analysis Plots**

*****Figure 5: Residual Plots of Balance, Income, Income^2, Gender, Cards and Rating*

**Residual Normality Test:**

****

****

****

*Figure 6: Normality plots for testing of Normal Assumption*