Modeling Average Credit Card Debt using Multiple Linear Regression

Anish K. and Albert(Xianzhi) W.

04/06/2020

Abstract

Motivation: Banks usually collect vast amounts of data on an individual when offering the service of a credit card. They use it to issue or reject credit card applications and set parameters such as credit limits. However, this same data can be used to predict the average amount of credit card debt held by an individual, which we do in this report.

Methods: Using data compiled on 400 individuals including several metrics, a trial and error methodology is used to produce a multiple linear regression model. The explanatory variables were picked out using various methods including stepwise regressions, best subsets, and the nested-F test.

Key Findings: The main predictors for average credit card debt of an individual include the individual's income earned, credit card limit, number of cards owned, age in years, and student status.

Concluding Remark: Average credit card debt given the small sample can be determined by a few key predictors, however, multicollinearity and the sample act as limitations.

Introduction

As we know already, there exists a lot of research undertaken by credit card issuers on determination of reliable borrowers. There exists a dearth of literature on one of the biggest issues the country and its individuals face: managing their credit card debt. After mortgages, auto loans and student loans this debt dominates loans owed by the American public. What drives this debt can help stem a potential crisis in the future. This study tries to anticipate the drivers using a sample of 400 individuals, 193 males and 207 females. The subjects average age is around 55 and the average years of education is 13 meaning they have on average completed high school. Further, 40 of them are students and more than 245 are married. They also have mean incomes of around $45,000 and average credit card debt of $520. The remainder of the paper seeks to explain how the data was modeled using preliminary analyses, testing various multiple regressions and transformations and finally arriving at an answer to research question: what attributes of an individual help predict their average credit card balance?

Method

After obtaining some summary statistics above we proceeded to first look at the outliers in the response variable credit card balance. Further, we also removed credit card balances that were zero. This was done due to a few reasons, one, because different kinds of people experienced zero balances regardless of the factors influencing them and secondly because it obscured the regression results and violated the assumptions of the OLS regression when trying to create single and multivariable regression models. This led to the elimination of around 90 zero balance values. Further, observing using a boxplot one more outlier was revealed which was subsequently eliminated. Now using the remaining 309 observations analysis could begin.

Data Analysis

We started the data analysis by firstly looking at the singe variable regressions of balance response variable and the other factors serving as explanatory variables. We first identify a handful of statistically significant factors by checking the p-values of their independent regressions on Balance and then proceed to construct a parsing model with variables being iteratively removed from the model based on the improvement in the AIC score using forward and backward stepwise regressions. Next we proceed to use the best subsets model to figure out the best multiple linear regression model

given the predictors. We then try and arbitrarily pick the best 3 predictor model and start with

statistical inference. Before beginning we first try and test out if the assumptions of the OLS regression are met. In order to do this we come up with at least 4 plots to investigate, the order versus residuals plot, the residuals versus the explanatory variable plot, the residuals versus fitted values plot and normal quantile plot. These test for the assumptions of independence of the residuals, no relationship between the residuals and the explanatory variables, constant variance of the residuals and the normal distribution the residuals respectively. From there as these assumptions are satisfied, the need for transformations to the data disappears. Moving onto statistical inference, we first obtain the confidence intervals for the slope coefficients in our model, followed by a two sample t-test for the Student explanatory variable and F-test for the same. We finally end up defining a full model and comparing our 3 predictor model with it. From the comparison it is revealed that more predictors are needed and we end up with out 5 predictor model. Again, we first make sure the assumptions of the model are satisfied and finish up with statistical inference from the model. We also estimate the need for an interaction term due to there being the Student categorical variable in our 5 predictor model. We end up not requiring a term from the interaction plots, since the non-student versus student lines with balance on the y-axis don't end up intersecting for both limit and income, this proves that there isn't a need for an interaction term. We also check to make sure multicollinearity is not an issue with this model, we end up with two predictors being highly correlated. We try several methods to see if we could eliminate the problem, by first eliminating one of the variables, and then by combining the two predictors and making it into one variable. Seeing that these methods reduce R^2 by more than 20% we decide to not take the action to remove multicollinearity.

Results

Starting of with the simple regression coefficients, the only ones with significant p-values for the slope coefficient t-tests are income, credit limit, credit rating, and student status and can be found in Table A in the Appendix. This means that individually their effects on the credit balance are noteworthy. This doesn't mean we should make a model with these, but they are strong predictors and will probably show up in our multiple regression models. From both the AIC score predicated forward and backward stepwise regressions we obtain a six predictor model including the four significant predictors from the simple regression case with age and number of cards as two others. The slope coefficients using the model are small for limit rating and age coefficients but larger for income and cards and the most for the student coefficient, so much so that if you are a student, credit card balance goes up by $425 dollars.

Table 1: Best Subsets Model

	Income	Limit	Cards	Age	Education	Gender	Student	Married	Ethnicity	Rating	.1
1	0	0	0	0	0	0	0	0	0	1	0.624
2	1	0	0	0	0	0	0	0	0	1	0.831
3	1	1	0	0	0	0	1	0	0	0	0.990
4	1	1	1	0	0	0	1	0	0	0	0.998
5	1	1	1	1	0	0	1	0	0	0	0.999
6	1	1	1	1	0	0	1	0	0	1	0.999
7	1	1	1	1	0	0	1	1	0	1	0.999
8	1	1	1	1	0	0	1	1	1	1	0.999
9	1	1	1	1	0	1	1	1	1	1	0.999
10	1	1	1	1	1	1	1	1	1	1	0.999

Table 1 above lists the results of the best subsets model that obtains the largest adjusted R^2 given a number of predictors. As can be seen after the 3 predictor model, the difference in adjusted R^2 doesn't improve by much. Thus we start inference with this three predictor model. As can be seen from Figure 1 - 4 in the appendix, the assumptions of OLS multiple regression have been preserved and therefore inference on this model can be conducted. Firstly to see if we should continue using this model we conduct a nested F-test with respect to the full model that contains all the predictors. The anova comes out with a significant p-value for the F-statistic meaning that there exist some non-zero predictors that are left out. When the anova is conducted for the 4 predictors from Table 1, again a significant p-value arises.

Table 2: Anova with 5 predictor Model

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
303	33722.24	NA	NA	NA	NA
297	32599.39	6	1122.842	1.704961	0.1194891

However when this test is conducted with the best 5 predictor model from Table 1 as can be seen in Table 2, we get a p-value above our significance level of 0.05 meaning that all predictors left out could have a value of 0, and thus more predictors shouldn't be added to the model. Thus, we arrive at a model with significant predictors from the F-statistic. This model is summarized in Table 3 below:

Table 3: Regression Results

	Dependent variable:
	Balance
	Main Effects
Constant	-702.743^{***} (-708.847, -696.639)
Income	-9.994^{***} (-10.052, -9.936)
Limit	0.327^{***} (0.325, 0.328)
Cards	24.839^{***} (24.012, 25.666)
Age	-0.997^{***} (-1.067, -0.927)
StudentYes	500.621^{***} (497.000, 504.242)
Observations	309
R^2	0.999
Adjusted R^2	0.999
Residual Std. Error	10.550 (df = 303)
F Statistic	$91{,}889.100^{***}$ (df = 5; 303)
Note:	*p<0.1; **p<0.05; ***p<0.01

Now the inference can be conducted again, by firstly making sure the assumptions of the OLS regression are met. As can be seen from Figures 5 - 8 in the Appendix, these assumptions are easily met and no transformations are required. In figure 5, we see the residual versus order plot being randomly distributed around 0, this means that the assumption about independence of residuals is satisfied. In figure 6, we see no relationship amongst the residuals and our explanatory variables from the randomly distributed points. In figure 7 we also see no relationship between fitted and residual values, meaning homoskedasticity is preserved and in figure 8, we see the residual points adhering to the normal probability line, meaning that the normality assumption is also preserved. Next, we move onto predicting whether we need an interaction term for the Student categorical variable. We can see clearly that we do not need an interaction term from Figure 9 and 10 in the Appendix. The figures shows that in the regression of Balance on Income and Limit respectively, the different lines for a student versus a non-student don't intersect, thus meaning there isn't a slope change in the regression due to the student variable. This result means that an interaction term won't be necessary.

Moving onto statistical inference we can see from the confidence interval results we obtain statistically significant coefficient values in Table B in the appendix. Since the intervals do not contain zero we can assume that all the slope coefficients in the model are statistically significant at the 5% level. Next, we examine the Student variable more and first look at the variance in Balance due to the Student explanatory variable using an F-test. The F-test yields an insignificant p-value of 0.067 at the 5% significance level and thus we have to assume that variances are unequal for the two groups when conducting a t-test. Next we conduct a two sample t-test for Student and the results are summarized in a Table C in the appendix. The p-value tells us that the difference in mean Balances for Students versus non-Students is statistically significant. These tests were conducted in the source code.

Lastly we need to make sure there exists no multicollinearity within the regression's explanatory variables. In order to do this, we first have to convert the Student categorical variable into a numeric type and we do this using the as numeric command. The correlation matrix in Table 4, below, shows us that there isn't much of a relation between the explanatory variables except for Limit and Income variables. The high coefficient of 0.83 might suggest effects of multicollinearity. However, in order to combat this a combination of the variables and an elimination of one of them was tested in the source code to no avail. The model's fit deteriorated tremendously and thus sticking with the issue seemed like the best option.

Table 4: Correlation Matrix

	Studentn	Income	Limit	Cards	Age
Studentn	1	-0.015	-0.127	-0.026	-0.021
Income	-0.015	1	0.826	-0.049	0.205
Limit	-0.127	0.826	1	-0.033	0.137
Cards	-0.026	-0.049	-0.033	1	0.015
Age	-0.021	0.205	0.137	0.015	1

Discussion

The interpretation of the results and the model we obtained can be interpreted as is but a few aspects need to be kept in mind. Firstly, the sample provided to us is just that a sample and does not reflect the population in any way, since the population is not just different but also constantly evolving as time progresses and millenials and generation Z accelerate consumer spending through credit card purchases.

In terms of the analysis, the decision to remove individuals who had a zero credit card balance proved to be critical because of the fact that when it remained, it detrimentally affected the predictions of the multiple regression model. The assumptions of constant variance of residuals and the normality of the residuals was blurred. This prevented any statistical analyses to take place with respect to any model derived. Even after transforming the dependent and explanatory variables, the residuals still didn't satisfy the normality or homoskedasticity assumptions. Thus it was essential to do so, since it discarded the confusing effects, for example someone with high income and low income could both be paying of their credit card debt on time regardless of their income, this results in similar credit card balance of 0, but different underlying factors.

Moreover, the determination of the model was also a bit construed since various methods were used in the process. The one that resulted in the final regression combined the best subsets technique along with the nested F-test. They help chose the appropriate predictors and their number respectively. All the slope coefficients in the model, however, indicate they are statistically significant as seen from the t-test p-values attached. Moreover, the F-test for the model when compared to an intercept only case also yields a highly significant p-value which means the model seems to be doing its job. However, we need to be careful when interpreting the final model, the model can be stated as follows:

$$\hat{Balance} = -702.74 + -9.94 * Income + 0.33 * Limit + 24.83 * Cards + -0.99 * Age + 500.62 * Student$$

The income coefficient here refers to what happens to the credit card balance if income rises by $1000 instead of $1. And the 500.62 is related to an increase in balance if the subject is a student. Further the two-sample t-test revealed that there is a significant difference in the mean Balance if you are a student or not, the 95% confidence interval is as follows: (-426.5, -105.9). The qualitative interpretation is that 0 isn't in this interval, this means that there is a significant difference in means of balance when one is a student versus when one is not. And this difference is large, being a student can mean a higher credit card balance by at least more than $105.9 and upto $426.5. Having more credit cards, interestingly increases credit card debt which can be intuitively understood since it would mean more spending as a whole. As age increases we see credit card debt decrease which can also be understood since older people tend to be more frugal, tend to save and fulfill

their debts. Finally as the credit card limit increases so does the debt, which also makes sense since the more the spending is allowed, the more people will look to take advantage.

The R^2 for the model is exceedingly high which seems more of a problem than being a positive quality. One of the possible explanations could be that multicollinearity could be causing an issue with the model that might be obscuring the true slope coefficients. Since, Income and Limit are so closely related and they are both included in the model, they must be having some distorting effects, since both should be technically unrelated. However, if one of the predictors is removed or if they are combined, the model greatly loses its predictive power as R^2 value drops to around 0.7. This creates a sticky situation with high multicollinearity obscuring the true slope coefficients.

Conclusion

Regardless of the limitations of our model, it has given us valid conclusions and a direction to pursue further research when looking at average credit card debt. The factors we identified as being crucial to affecting credit card debt included again the individuals income, credit limit, student status, credit cards owned, and their age. Their interpretations also make sense. While the student status did not have an interaction effect, it has a large impact on the credit card debt and shifts the intercept by more than $500 above if the individuals turns out to be a student. This tells us that the student demographic need special consideration when looking at credit card debt of individuals and less so the elderly and those with lower credit limits. A study with perhaps more of a younger target demographic and a more representative sample could make the study even stronger. Nevertheless, multicollinearity continues to remain an issue due to the obvious relationship between the income and limit predictors.

References

- Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

- Cannon, Ann R., George W. Cobb, Bradley A. Hartlaub, Julie M. Legler, Robin H. Lock, Thomas L. Moore, Allan J. Rossman, and Jeffrey A. Witmer. STAT2: Building Models for a World of Data. New York: W.H. Freeman, 2013.

Appendix

Table A: Simple Regression Output

Variable	R2	Slope_Coef	P_value
Income	0.154	4.30	0.000000
Limit	0.621	0.16	0.000000
Rating	0.625	2.40	0.000000
Student	0.047	266.22	0.000118

Table B: Confidence Intervals for Slope Coefficients

	2.5 %	97.5 %
(Intercept)	-708.8711392	-696.6149653
Income	-10.0521530	-9.9359924
Limit	0.3254809	0.3276309
Cards	24.0091106	25.6691694
Age	-1.0677044	-0.9267987
StudentYes	496.9858911	504.2569131

Table C: Two-sample t-test with unequal variances

	unlist.table.
statistic.t	-3.34355802654616
parameter.df	45.5539743657139
p.value	0.00166146988297608
conf.int1	-426.530745688429
conf.int2	-105.906861149177
estimate.mean in group No	633.088888888889
estimate.mean in group Yes	899.307692307692
null.value.difference in means	0
alternative	two.sided
method	Welch Two Sample t-test
data.name	Balance by Student

Figure 1: Order plot. Assumption of independent error terms is preserved from order plot

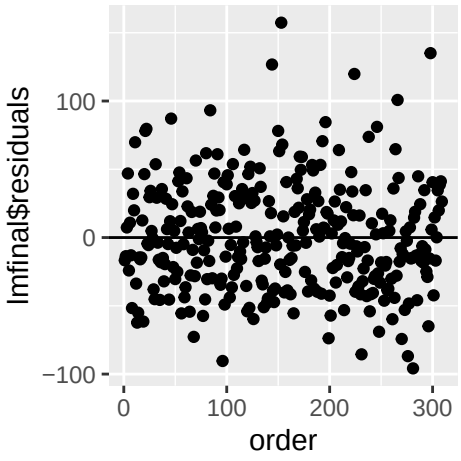

Figure 2: Plots with explanatory variables. Shows assumption of independent explanatory variables and residuals is preserved

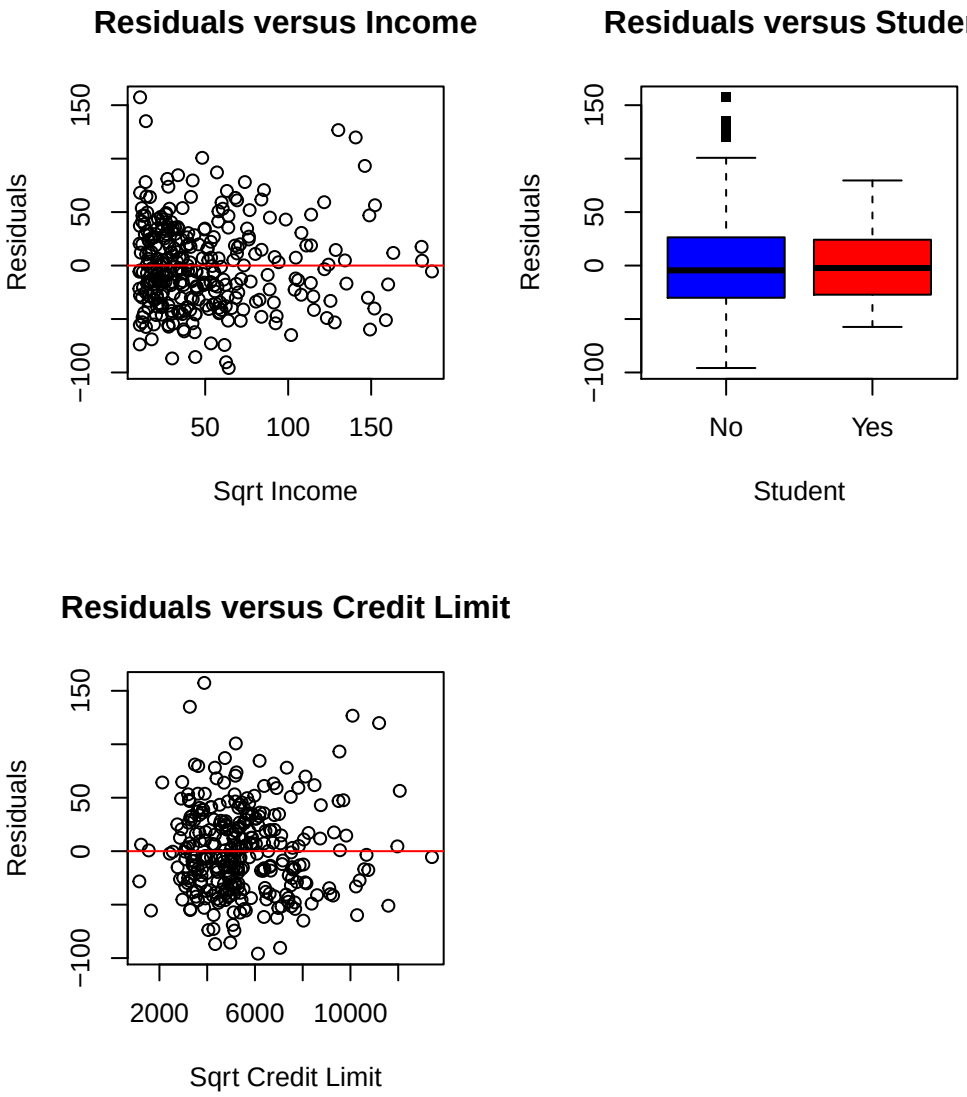

Figure 3: Residuals vs Fitted plot. Shows assumption of constant variance/homoskedasticity is preserved

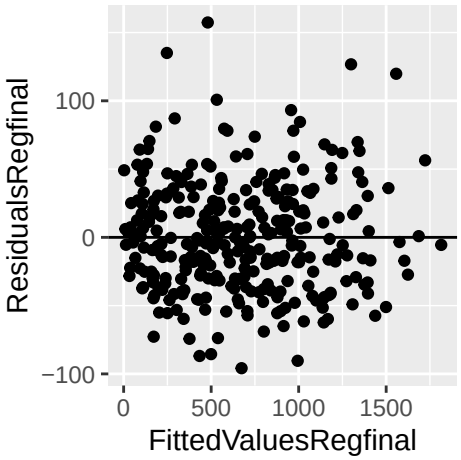

Figure 4: Normal probability plot. Shows how error terms are normally distributed.

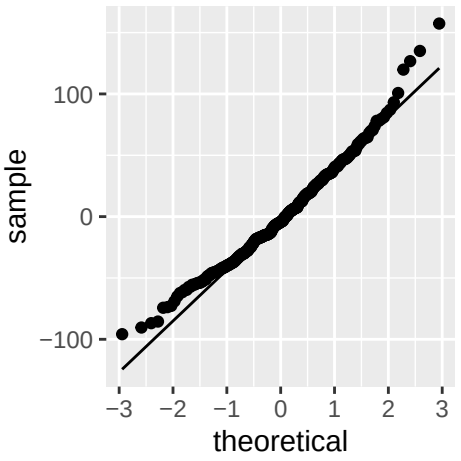

Figure 5: Order plot - 5 predictor model

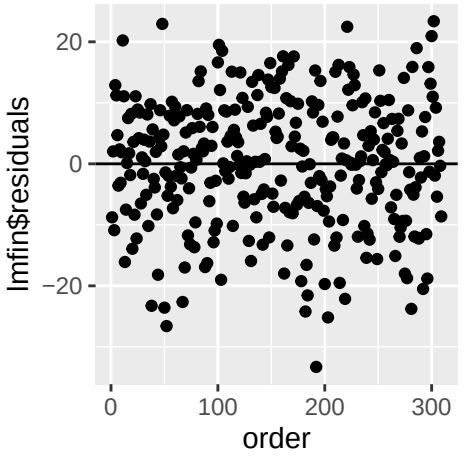

Figure 6: Plots with explanatory variables - 5 predictor model

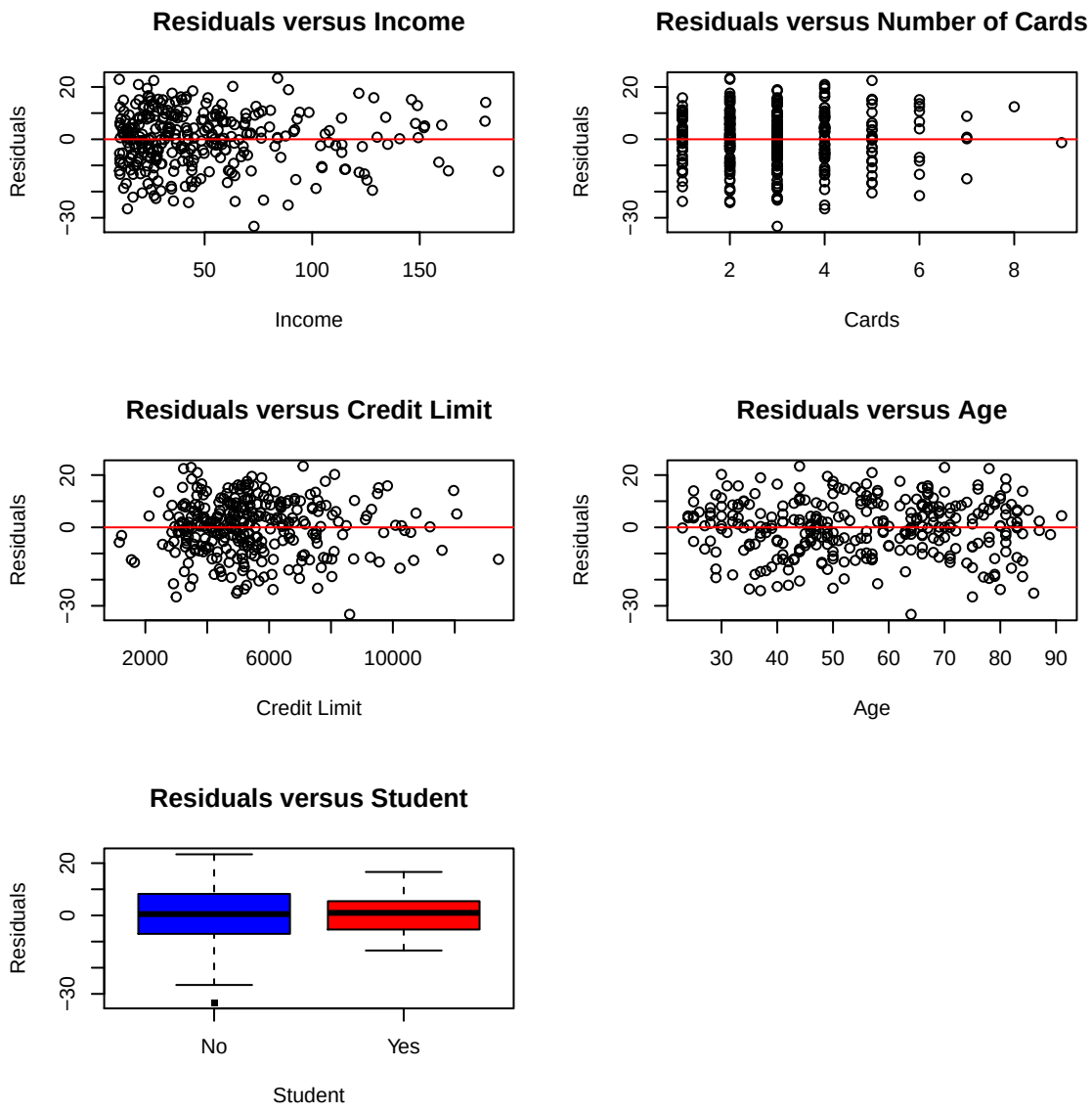

Figure 7: Residuals vs Fitted Plot - 5 predictor model

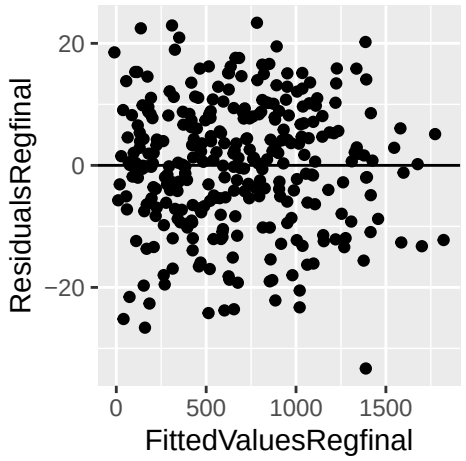

Figure 8: Normal Probability Plot - 5 predictor model

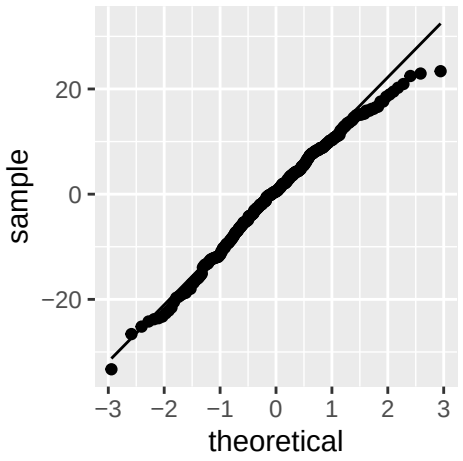

Figure 9: Interaction Term - Limit

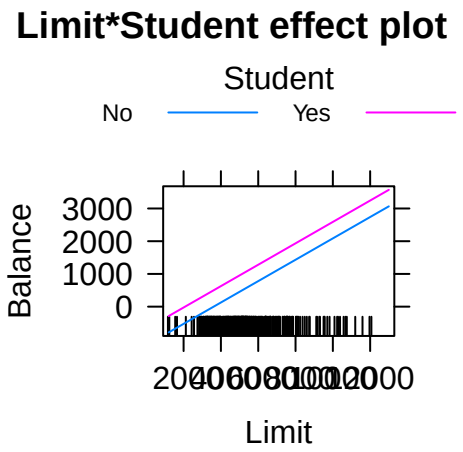

Figure 10: Interaction Term - Income