Credit Card Analysis

Gabriel Adams

January 2019

1 Introduction

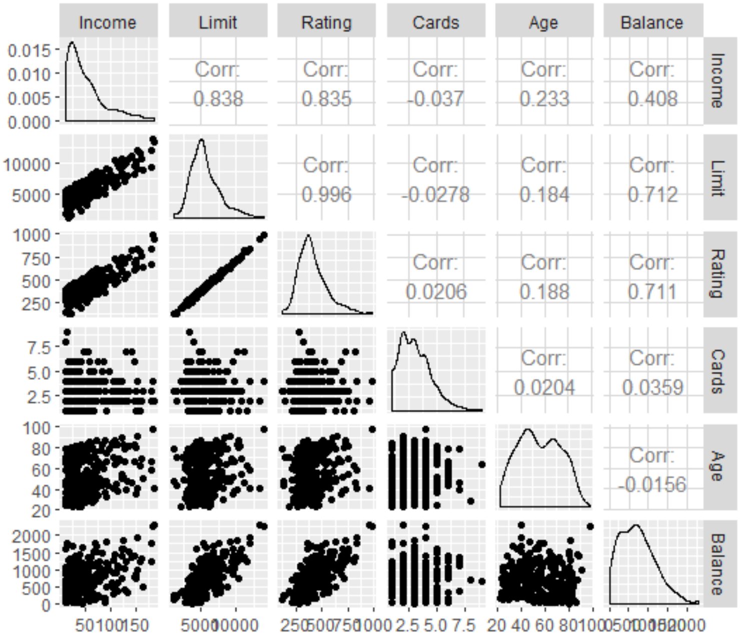
Interest paid on credit cards is the largest source of profit for credit card companies. The size of the balance carried by an individual provides quite a bit of information concerning the potential profit that individual will provide to the company. Individuals with low balances are low profit, but also have a low risk at defaulting. Individuals who default on their credit card debt are a large risk for credit card companies because credit cards are unsecured and it can be quite costly, even infeasible, to recover lost funds from customers who have defaulted. Individuals with high balances, generally, are a greater risk to the company because of their high risk of default. Individuals with moderate balances provide the ideal situation for credit card companies because they provide an ideal balance of risk and profit. Because of this, credit card companies are interested in predicting credit card balances ahead of time. With this information, these companies will be better equipped to make decisions on extending credit cards to customers.

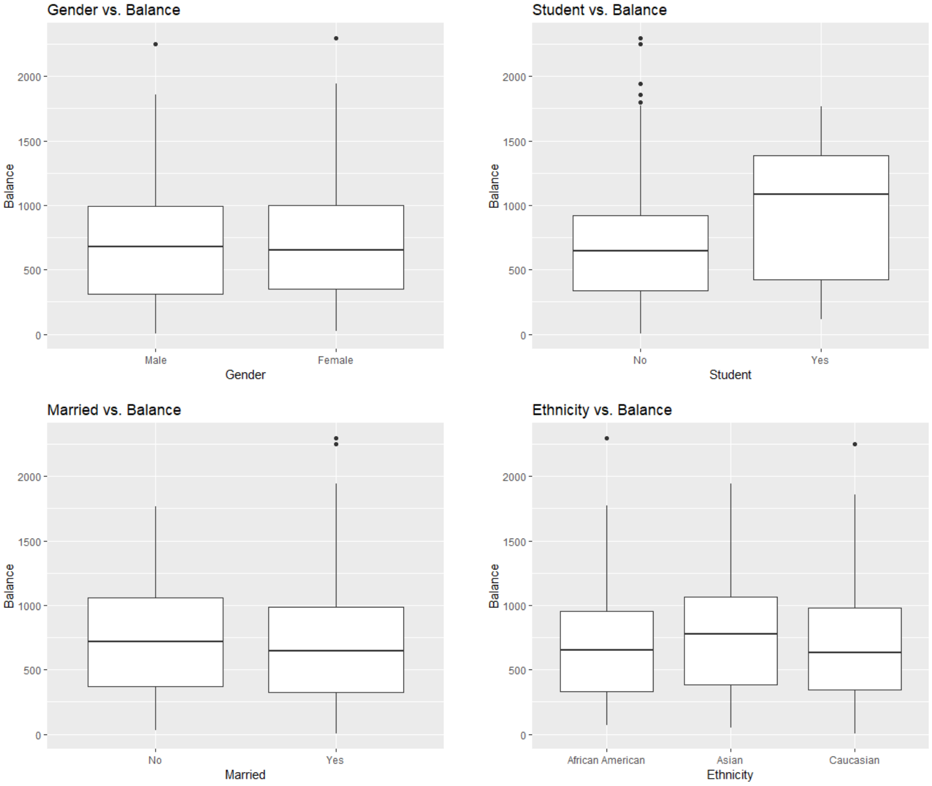
Our primary goal is to accurately predict card balances based on several known factors of potential consumers. These predictions, in turn, will aid credit card companies in their decisions on whom to lend to. We are also interested if people become more responsible with money (lower balances) as they grow older. Our final piece of inquiry is to research the efficacy of a credit card company's policy of increasing credit limits of customers whose incomes increase (the credit limit is increased at a rate of 10% of the income increase).

To accomplish our research goals, we will use a dataset from a certain credit card company consisting of data from 294 credit card holders. This dataset contains their balance and several explanatory variables. These explanatory variables are income (measured in thousands of dollars), credit limit, credit rating, the number of cards held by the holder, age, education status, gender, student status, marriage status, and ethnicity.

The first plot on the following page is a pairs plot which compares all of the quantitative variables with each other. We can see scatterplots and correlations. On the bottom row, we can see that Income, Limit, and Rating appear to have positive linear relationships with Balance. It is important to note, however, that Income, Limit, and Rating are all highly correlated with each other. They are collinear. Basically, if we know one of the variables, the other variables do not give us much new information. When we fit our statistical model later, we will have to take this into account and perhaps remove one or two of these variables from our final model.

Below the pairs plot are four boxplots, one for each of the categorical variables. Gender, ethnicity, and marriage status do not seem to have much of an effect on Balance, while students do seem to have a higher balance than non-students. This makes intuitive sense. We will address these relationships more formally later in the report.





2 Model Selection

We will fit an appropriate statistical model to the data and use this model to predict balances and answer our other research questions. This model will use some or all of the previously mentioned explanatory variables.

A multiple linear regression model will be an appropriate tool to analyze our data and answer our questions. We use variable selection to avoid, as best we can, underfitting and overfitting. Variable selection will also help us solve the problem of collinearity, which was mentioned previously. Because we have relatively few explanatory variables and are primarily interested in prediction (predicting balances), we use an exhaustive method with AIC. This method will give us the best set of explanatory variables to use in our model. Running this selection algorithm gives us a model with five explanatory variables: Income, Limit, Cards, Age, and Student.

Our model is as follows:

$$\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$$

- $\boldsymbol{y}$ is a vector of all the averaged final scores (our response variable).
- $\boldsymbol{X}$ is the design matrix containing relevant information from our observations. Its first column is a column of 1's, corresponding to the intercept. The second through tenth columns are correspond to the second through tenth semesters (the first semester is accounted for in the intercept). In each column, a 1 for a specific observation signifies that the course belonged to that semester. A 0 in a column means that the course did not belong to that semester. If there are 0's in all the columns between the second and tenth, then the course was in the first semester. The next three column are vectors of the average scores for the three midterm exams, with one entry in each vector corresponding to a specific course. In a similar fashion, the last two columns in the matrix correspond to the average homework grades and quiz scores.
- $\boldsymbol{\beta}$ is a vector of coefficients. The first element of $\boldsymbol{\beta}$ (β_0) is the intercept term. Thus, when all scores are 0 and a course is in the first semester, the predicted balance will be β_0.
- $\sigma^2\boldsymbol{i}$) indicates that the residuals have an equal variance about the line and that our observations are independently distributed (0's in the off-diagonal). Our assumptions of equal variance and independence are at play here. We will verify these assumptions in the next section.

This model serves our purposes well because it will allow us to, by plugging in known information, make predictions about future balances. We will also be able to calculate the effects specific factors have on Balance. This will help us determine if people become more fiscally responsible as they age. This model will also help us analyze the credit card company's policy of increasing credit limits by 10% of a corresponding increase in income.

Assuming our assumptions hold true, this model will ultimately prove quite useful in our analysis. These assumptions are linearity, independence, normality, and equal variance. We will verify these assumptions in the next section.

3 Model Justification and Performance Analysis

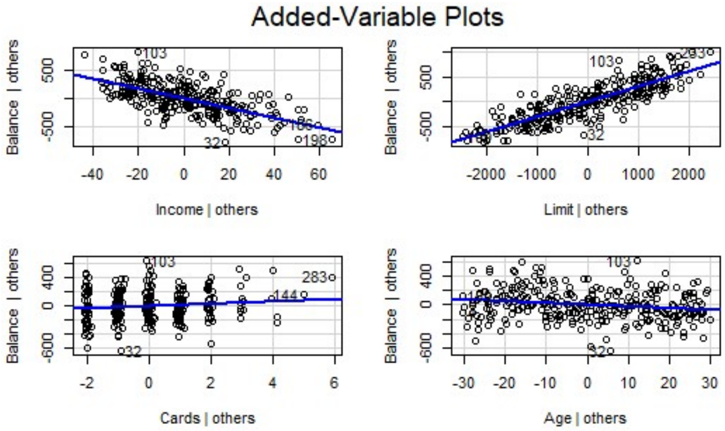
Again, to fit our model, we used best subset selection with AIC. One of our main goals is predicting balances, so we chose AIC (which emphasizes prediction). We have a relatively low number of explanatory variables, so best subset selection was appropriate and computationally feasible.

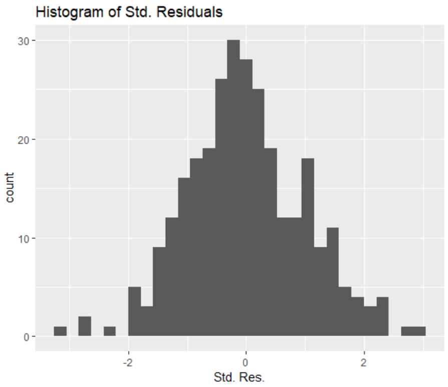
To properly use multiple linear regression models, four assumptions must be verified: linearity, independence, normality, and equal variance. We now verify each assumption.

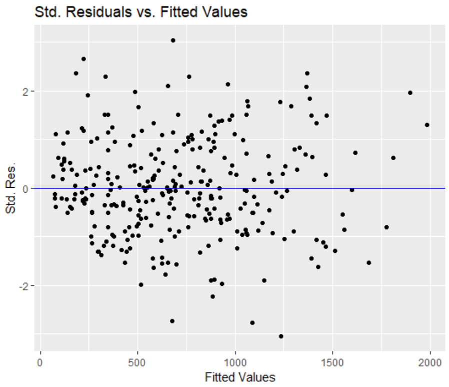
1. Linearity can be verified through the use of added-variable plots. In the added-variable plots on the next page, we can clearly see linear trends in the quantitative variables. Nothing looks too nonlinear. Thus, our assumption of linearity holds.
2. It is reasonable to assume that the balance of one person does not affect the balance of another. This means that we can assume independence.
3. Normality can be verified through a histogram of standardized residuals (below left). The standardized residuals appear to follow a standard normal distribution. As an additional check for normality, we perform a KS test in R. This test has a null hypothesis that the standardized residuals are distributed according to the standard normal distribution. Running the test in R gives us a p-value of 0.5755.

Because this p-value is greater than 0.05, our chosen alpha value, we fail to reject the null hypothesis and conclude that the standardized residuals are normally distributed. Our assumption of normality holds.

4. Equal variance can be verified by looking at a plot of standardized residuals vs. fitted values (below right). The spread of the points about the fitted line is roughly the same as we move from left to right, so equal variance is reasonable to assume. The BP test has a null hypothesis that there is equal variance. Running this test in R returns a p-value of 0.0615. This p-value is only slightly greater than 0.05, so it is a borderline case. However, by looking at the plot, we can reasonably conclude that there is equal variance.



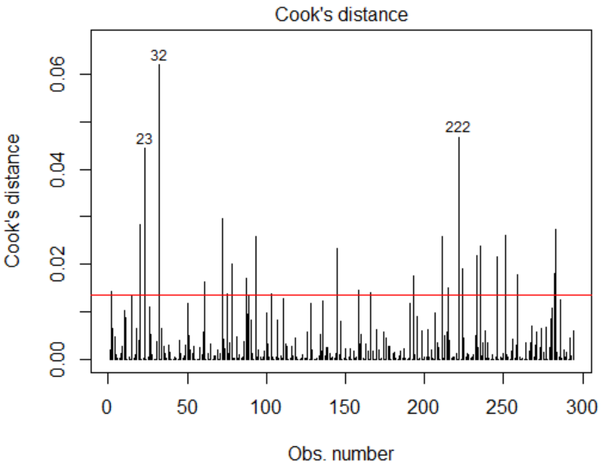




Our four assumptions hold. It does not appear that a transformation is necessary for this data. However, there are few outliers in the data that could be concerning. Calculating Cook's distance and comparing these values to the rule of thumb shows that 26 observations are of concern. We can seem them in the plot on the next page.

To see if these outliers make a meaningful difference in our estimates, we remove the three most egregious outliers (23, 32, and 222) and refit our model using the same explanatory variables. It turns out our new estimates for $\boldsymbol{\beta}$ are about the same as before, so we will basically ignore the outlier issue by keeping them in our data. This will not affect our analysis in any meaningful way.

We were also informed of a potential interaction between the student and income variables. To test for this interaction, we run an F test in R. Our null hypothesis is that the beta coefficient associated with this interaction term is zero. This test returns a p-value of 0.8769. Because our p-value is greater than 0.05, we fail to reject the null hypothesis and we conclude that the interaction is not significant. Thus, we will not add an interaction term to our model.



In R, we calculate R^2. This model has an R^2 of 0.7845. This means that 78.45% of the variation in balance is explained by our model (the selected explanatory variables). This means our model fits the data well.

We are interested in how well our model predicts. To assess our model's predictive capabilities, we run a Monte Carlo cross validation study in R. We calculate an average bias value of -1.79 and an average RMSPE value of 214.77. This means that, on average, our model will predict values for balance $1.79 below their actual values, and that each prediction, on average, will be $214.77 "away" from the actual value. On the scale of our data, these values are reasonable and are not concerning.

We are not only interested in the accuracy of our point predictions, but our prediction intervals as well. Using R, we determine that, on average, our 95% prediction intervals have a width of $852.97. This means that we can typically predict within $\frac{853}{2} = 426.5$ dollars with 95% confidence. On the scale of our data, this is pretty good. Using R, we can also see that our coverage is approximately 94%. This means that we should treat our 95% prediction intervals more like 94% prediction intervals because our model is producing intervals which only contain the actual values 94% of the time. 95% and 94% are close enough that this discrepancy should not be cause for worry.

Overall, it appears that our assumptions hold and that our model fits well and has good predictive abilities. We will use this model to analyze the data and find answers to our research questions.

4 Results

The maximum likelihood estimates (calculated in R) for individual elements of $\boldsymbol{\beta}$ are in the following table. Each estimate is accompanied by a 95% confidence interval, which allows us to see a measure of uncertainty in our estimates.

	Effect	Estimate	Lower	Upper
β_0	Intercept/Baseline	-547.75	-671.87	-423.63
β_1	Income	-8.51	-9.69	-7.32
β_2	Limit	0.31	0.28	0.33
β_3	Cards	16.59	-0.21	33.39
β_4	Age	-2.34	-3.77	-0.90
β_5	Student (Yes)	531.83	457.10	606.56

Some of the results in the table above are somewhat unexpected, but, after some thought, make sense. A hypothetical non-student who is zero years old with zero income, no credit limit, and no cards would have, on average, -$547.75 as a balance, with a 95% confidence interval from -$671.87 to -$423.63. This hypothetical situation is absurd, so we will continue with our analysis. Holding all else constant, we are 95% confident that a person whose income increases by 1 unit ($1000) will have their balance decrease by between \$7.32 and \$9.69, with a best guess of \$8.51. This does not make much intuitive sense, but we should remember that a lot of the same information contained in Income will be captured by the variable Limit. More importantly, Income is on a different scale as Balance so this affects our intuition. Holding all else constant, we are 95% confident that a person whose limit increases by \$1 will have their balance increase by between \$0.28 and \$0.33, with a best guess of \$0.31. Holding all else constant, we are 95% confident that a person who gains an additional card will have their balance change by between -\$0.21 and \$33.39, with a best guess of \$16.59. This variable is borderline significant, so that could explain the small part of the confidence interval that is negative. Holding all else constant, we are 95% confident that a person who gains a year will have their balance decrease by between \$0.90 and \$3.77, with a best guess of \$2.34. Finally, we are 95% confident that a student will have a balance of between \$457.10 and \$606.56 greater than an otherwise equivalent non-student, with a best guess of \$531.83.

We can now begin to answer our research questions. As shown in the previous section, our model is quite good at predicting balances given we know the values for various explanatory variables. Credit card companies can use this model to predict estimates of future balances, as well as calculate prediction intervals to quantify uncertainty.

From our table above, we can conclude that as people age, their balances do tend to decrease. Again, holding all else constant, we are 95% confident that a person who gains a year will have their balance decrease by between \$0.90 and \$3.77, with a best guess of \$2.34.

We are also interested in evaluating the credit card company's policy of increasing limits by 10% of income increases. The expected difference in balance when income goes up by 10% is approximately \$221, with a 95% confidence interval from \$206.51 to \$234.26. That is, when someone's income increases by \$10,000 and the company follows its credit increase policy, we are 95% confident that Balance will increase by between \$206.51 and \$234.26. Overall, the company's policy seems reasonable. An automatic limit increase could be unwise, however, if there are other factors showing that a customer is more likely to default. Some of these factors, like total credit extended across all accounts, could potentially be included in a model to help prevent default and loss of profits for the company. Overall, however, the policy seems reasonable.

5 Conclusions

Through use of a justifiable multiple linear regression model, we were able to answer the questions posed at the beginning of the analysis. Using the known variables Income, Limit, Cards, Age, and whether or not someone is a student, credit card companies can accurately predict balances within a certain interval. This predictive capability will allow credit card companies to better manage their risk and gain more profits.

We also discovered that as people age, their financial responsibility generally increases as well (lower balances). This makes intuitive sense and could be a useful piece of information for credit card companies.

We also determined that the expected difference in balance for someone whose income increases by \$10,000 is approximately \$221. The policy of the credit card company to increase credit limits by 10% of an income increase seems reasonable, but could be improved by adding additional safeguards such as taking into account total credit extended across all accounts.

Our model has several potential weaknesses and shortcomings. We are given limited information on cardholders. Although our model fits the data well, there might be other variables out there that more thoroughly explain balances. For future research, it would be beneficial to collect other explanatory variables such as zip code and other types of loans/credit to better predict balance. Type of employment could also be useful. Finally, there is certainly a more clever way of determining how to extend credit limit increases. This is worth thinking about more.