Case Study 1
Stat 536 - Dr. Heaton

Jacob Mortensen & Lot Slade

January 22, 2015

1 Introduction

Like most businesses, credit card companies are striving to maximize profits, and the best way for a credit card company to accomplish this objective is by having customers who keep a moderate balance on their credit card. Unfortunately not all customers exhibit such behavior, some have very low balances and are not paying interest, while others have very high balances that may or may not ever be repaid. In both cases the credit card company isn't making very much money, but what if a company could predict the level of balance a customer would have before ever approving them for a card? The ability to predict a potential customer's balance would allow a company to fill their customer pool with only those persons who have a higher probability of making them money.

The goal of this analysis is just that, to predict the credit card balance of potential customers. This will be done using a data set of 400 observations collected on the following variables:

Variable Name	**Description**
Income	Card holders monthly income in thousands.
Limit	Card holders credit limit.
Rating	Credit rating - similar to a FICO credit score but used internally by the company.
Cards	Number of open credit cards (including the current card) of the card holder.
Age	Age of the card holder.
Education	Years of education completed by the card holder.
Gender	Gender of the card holder.
Student	Card holder is a full-time student.
Married	Card holder is married.
Ethnicity	Card holder's ethnicity.
Balance	Current credit card debt

After an initial exploratory data analysis, it was found that the data are not normally distributed and that Rating and Limit are highly associated with each other.

2 Methods

In order to properly analyze this data we elected to use multiple linear regression. The reason for using this method is that linear regression can be a powerful tool when you are attempting to predict an outcome given a limited number of inputs. In this context, creating an accurate linear model would allows us to predict the probable account balance of a future customer. However, in order for us to use this method there are a few assumptions that have to be met. First, that the relationship between the variables is linear, meaning that the effect of each variable can be modeled by a straight line. Second, that the errors have constant variance and that as the response variable increases in value, the variance of the residuals neither increase nor decreases. Third, that the errors are independent, which simply means that the size of one error has no impact on the size of another. Fourth, that the errors are normally distributed. We will discuss how well our model meets these assumptions in the model justification section.

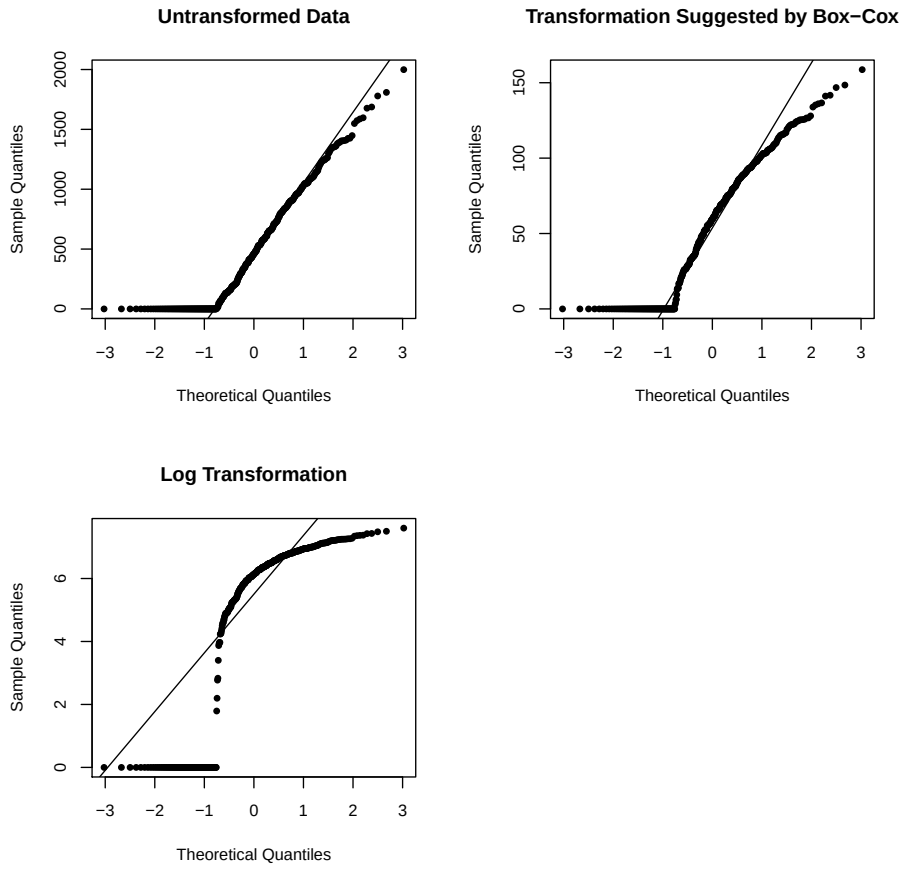

Figure 1: QQ Plots demonstrating normality in the Balance variable under various transformations.

We began by splitting our data into a training and test data set so that we could develop our model using one set of data and validate its accuracy using the other. We excluded the limit variable from our data set because our exploratory data analysis showed that it had a high degree of collinearity with the rating variable. We are concerned about collinearity because if two variables are highly collinear then small changes in the data could have a big impact on our estimates of the model coefficients. Additionally, at an intuitive level it makes sense that the credit card company would use a customer's rating to determine their credit card limit, and so limit and rating essentially act as a measure of the same thing. In our exploratory data analysis we generated some QQ plots and found that our response variable, Balance, exhibited extreme skewness. We attempted to mitigate this by using a transformation but found that our proposed transformations actually made this issue worse, both through visual inspection of the plots in Figure 1 and by conducting the Shapiro-Wilk normality test on the transformed and untransformed response. For these reasons we decided to continue to use the untransformed response despite the obvious issue of it having a preponderance of zeroes and being heavily skewed.

After addressing these concerns, and using a process that will be further described in the model justification section, the model that we decided on is

$$\begin{aligned}\text{Balance } &= \beta_0 + \beta_1 * \text{Income } + \beta_2 * \text{Rating } + \beta_3 * \text{Student} \\ &\quad + \beta_4 * \text{Income*Rating } + \beta_5 * \text{Student*Rating} + \epsilon\end{aligned}$$

where β_0 is the intercept, β_{1-7} are the coefficients corresponding with each variable, and ϵ represents the error term (i.e., the variation in Balance not accounted for by our covariates).

3 Model Justification

We conducted our initial variable selection using a forward stepwise selection technique in R. Because our exploratory data analysis had suggested that some interaction effects may exist, we opted to include all interactions as possible candidates for inclusion in the model in addition to the main effects. This left us with 67 variables to test, making best subset selection computationally intensive, which is why we opted to use forward selection. We decided to use the model with the lowest Bayesian Information Criterion (BIC) as our starting point for our model selection because it has a fairly large penalty for each variable added, ensuring that we include only the most significant variables, allowing us to develop a powerful model while keeping it relatively simple. The model suggested by this approach was the following:

$$\begin{aligned}\text{Balance} = \beta_0 &+ \beta_1 * \text{Income } + \beta_2 * \text{Rating } + \beta_3 * \text{Student } + \beta_4 * \text{Income*Rating} \\ &+ \beta_5 * \text{Income*Age } + \beta_6 * \text{Rating*Student}\end{aligned}$$

We further confirmed that this was a good model by calculating the MSE generated by this model on the test data set and found that of all the models suggested by the forward selection method, this particular one yielded the lowest MSE on the test data set.

Using the above model as a starting point we added Age due to the inclusion of Income*Age and fit it in order to examine its performance. We found that the

estimated effect of the interaction between Income and Age, while statistically significant, was practically negligible and that excluding that interaction and the Age variable decreased the adjusted R^2 from 0.9567 to 0.9553, a difference of only 0.0014. We also found that when validating the model with the test data that our coverage actually improved when we excluded age, moving from 93.7% of observed Balances in the test data set falling within the prediction interval to 96.0% falling within the prediction interval (and just to satisfy curiosity, the average width of the prediction interval was nearly the same in both cases, a range of about $192.50 to either side of the point estimate). Although neither value is exactly the 95% coverage we expect our prediction interval to obtain, because our goal is prediction, we decided that is better to err on the side of getting a few more correct predictions rather than a few less. Additionally, removing Age and Income*Age from our model makes it easier to interpret and where these variables are providing us with very little additional information about Balance we feel justified in removing these two terms from our model.

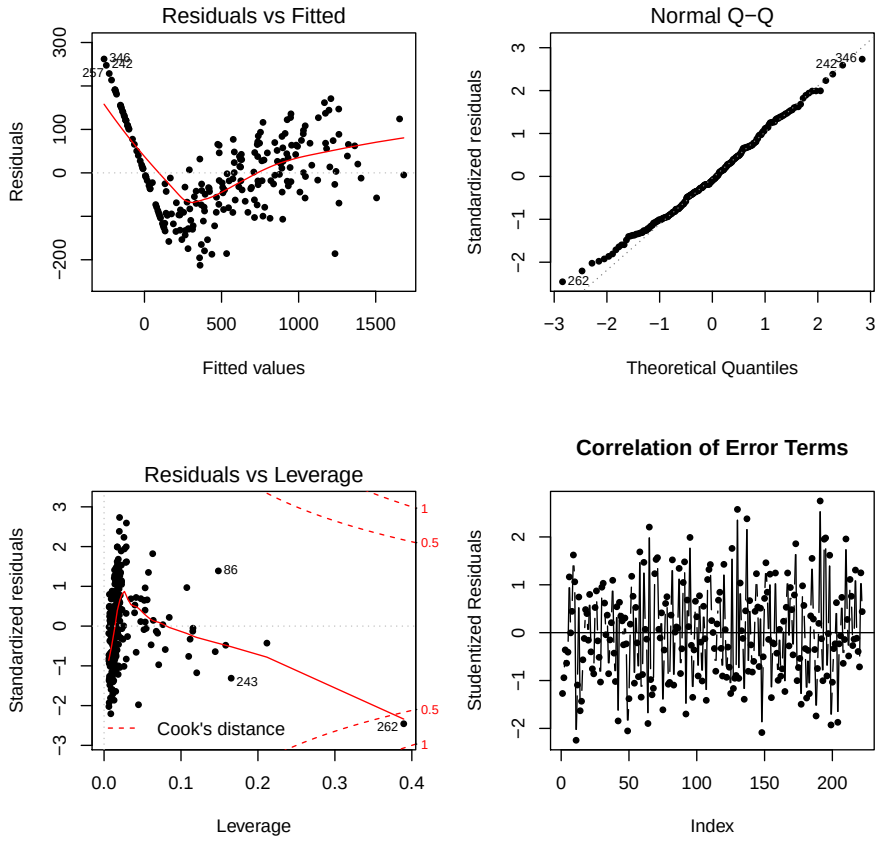

Figure 2: Plots allowing us to examine the assumptions of linear regression

Here we also examine the four main assumptions of linear regression: (1) that the relationship between the variables is linear, (2) that the errors have

constant variance, (3) that the errors are independent from one another, and (4) that the errors are normally distributed. If we examine Figure 2 then we can see how well each of these assumptions is met.

In the Residuals vs. Fitted plot in the upper left corner we can see that the smooth line is not perfectly horizontal which might normally indicate a non-linear relationship, but as we expected already, this is probably due to the large number of zeroes in the response variable, and we have been unable to successfully provide a preferable alternative to leaving the data as it is. Other than the pattern due to the large number of zeroes, we see little else in this plot to indicate that the relationship between the variables is not linear. We can also see from this plot that our second assumption, that the variance of the errors is constant, appears to be met once you account for the skewness of the data, as most of the residuals are a consistent distance from the smooth line all along the x-axis. The correlation of error terms plot shows that there are no visible trends in the errors, and so we conclude that our assumption that the errors are independent is met. Our QQ plot demonstrates that the fourth assumption, that the errors are normally distributed, appears to be met as well, and considering that we know our data is heavily skewed our plot actually looks quite good.

One other thing that is not necessarily related to one of the four main assumptions, but is nonetheless important, is influential points. We can see from the Residuals vs. Leverage plot that there are several points with higher leverage, and one point in particular. In order to see what kind of an effect they had on the model we removed them from the data set and refit the model. What we found was that while the size of some of the coefficients changed a little bit in some cases, overall things were close to the same, and the level of significance did not change in any of the cases, so we are comfortable leaving those points in the data set and keeping the model as it is.

Of course, because we are interested in prediction, the real determining factor in whether or not our model is justified is how well it performs in predicting the balance of certain individuals. As previously mentioned, when we cross-validated our model using the test data set we found that 96.0% of the observed Balances in our test data set fell within the 95% prediction interval calculated using our model. Therefore, we feel that our model fits the data well and is appropriate to use for the purposes of predicting an individual users credit card balance.

4 Results

The table below shows the estimated effect of each of the variables included in the model.

Variable	**Estimate**	**95% CI lower**	**95% CI upper**
Intercept	-476.60	-531.90	-421.32
Income	-10.84	-12.17	-9.50
Rating	3.79	3.62	3.95
Student	290.10	173.74	406.37
Income*Rating	0.0051	0.0030	0.0073
Student*Rating	0.3668	0.0476	0.6860

Following are contextual interpretations of each estimate.
Intercept: The average balance on a customer's card given Rating=0, Income=0 and Student=No.
Income: For every $1,000 increase in Income, Balance decreases by $10.84.
Rating: For every 1 unit increase in Rating, Balance increases by $3.79.
Student: If a customer is a student, Balance increases by $290.10.
Income*Rating: For every 1 unit increase in the interactive effect of Income and Rating, Balance increases by a half of 1 cent.
Student*Rating: For every 1 unit increase in the interactive effect of Student and Rating, Balance increases by $0.37.

As can be seen, with a fairly small amount of information an accurate prediction of Balance can be made. In fact, it has been shown that the model containing the collection of variables described predicts future balances with approximately 95% accuracy. That is, in the long run the prediction intervals produced from this model will contain the true balance 95 times out of 100.

5 Conclusion

By simply inputting information about a potential customer into the model, a predicted average balance can be calculated. With a better understanding of what type of balance that customer will carry in the future, the company now has the ability to screen out those customers who are more likely to not repay debts while accepting those applicants who will keep a moderate balance and bring in revenue.

While we feel that our model performs accurately, there remains the possibility for improvement meeting the normality assumption and slimming down the prediction intervals. As mentioned earlier, we recognize that Balance is not distributed normally, which is one of the assumptions involved in multiple linear regression. Potential transformations were explored and found to not be very helpful, but other transformations could be explored and implemented to meet the normality assumption. Another aspect of our model that could possibly be improved is the width of the prediction intervals. Currently the average width of the prediction intervals is $385. As they are, the prediction intervals are still useful but are a bit wide; slimming these intervals would give even better prediction than those that are currently being produced.

The next steps in this analysis would be to address these potential improvements. If we could find a way to effectively deal with the violation of our normality assumption it is possible that our prediction intervals could become much narrower without losing any coverage. Additionally, as we performed our analysis we saw evidence to suggest that a model containing only three variables (Income, Rating, and Student) may perform nearly as well as the one we decided on. It would be interesting for future analysis to see if we could develop an even simpler model that still performed adequately in predicting future balance.