Credit Card Analysis

David Lowe

January 31, 2017

Abstract

Credit card companies make their money by charging large amounts of interest on money loaned to their clients. Clients with moderate to high balances prove profitable to the company, as long as their debt does not overwhelm them and lead to loan defaults. It is of great interest to credit card companies to know which clients attain the optimum balance, and to be able to predict the balance of future potential clients. This study is to create a statistical model that can demonstrate the relationship between credit card balance, and characteristics of the credit card holder. This data was taken from internal data from a credit card company. We will be exploring the use of multiple linear regression to answer our desired questions.

1 Introduction & Data

Credit is based on the idea that money borrowed should be paid back with an added amount of interest, due to the creditor's risk in lending the funds. Credit card companies use this principle by lending funds to their clients that does not accrue interest under a given time, hoping that they will be unable to pay the full amount, and will rely on making monthly, interest accruing payments. This, however, requires credit card companies to maintain a delicate balance in their clientele. Those customers that pay their balance before accruing interest provide little to no benefit to the credit card company. Those who have high credit card balances may fold under the weight of debt, and file for bankruptcy, leaving the credit card company without recompense for their loans. The best clients are those those that have a moderate amount of credit card balance, allowing the credit card company to charge interest, but not so much that the client fails to pay.

Understanding what kind of client will produce these medium range balances becomes a point of interest for credit card companies. They often gather data on their clients (and future clients) to make more informed decisions on who to allow to use their credit cards. In this dataset, information such as age, ethnicity, education, and credit limit and credit rating were collected by a credit card company, along with the outstanding monthly balance. From this information they hope to be able to predict a future client's balance, but also understand the relationship these variables have with credit card balance, for marketing decisions and other internal use.

This analysis will primarily focus on creating a model that can predict, with accuracy, the amount of balance a future client will have. Such a model will allow the credit card company to predict who will be the most profitable clients, and thereby improve the selection process and boost revenue. Since we are testing the effect of several characteristics on credit balance, a multiple linear regression (MLR) model seems to be appropriate, since it is primarily focused on measuring the strength, and nature of the relationship between explanatory variables and the response (monthly credit balance in our case).

The data is structured with several numeric variables, some continuous and some discrete, each constrained by their own bounds. Our dataset also contains categorical variables with two or more levels. The table shown below describes the provided data, and defines the space the variables span. Some variables show high collinearity, such as with credit card limit, and credit rating. These variables can increase the variability of the coefficients. We must account for this in our model selection. The relationships between each numeric variable seem to be best described linearly, and the categorical variables provide a type of linearity by moving from one level to another. This strengthens the argument to use multiple linear regression to map this relationship.

Variable	Description	Type	Support/Levels
Balance	Current credit card debt	Continuous	$[0, \infty)$
Income	Monthly income (000's)	Continuous	$[0, \infty)$
Limit	Credit limit	Continuous	$[0, \infty)$
Rating	Internal Credit Rating	Continuous	$[0, \infty)$
Cards	# of credit cards	Discrete	$0, 1, 2, ...$
Age	Age of card holder	Discrete	$0, 1, 2, ...$
Education	Years of education	Continuous	$[0, \infty)$
Gender	Gender of card holder	Categorical	Male, Female
Student	Card holder is a full-time student	Categorical	Yes, No
Married	Card holder is married	Categorical	Yes, No
Ethnicity	Card holder's ethnicity	Categorical	African American, Caucasian, Asian

2 Method & Model

2.1 Model

$$Balance_i = \beta_0 + \beta_1(Income_i) + \beta_2(Limit_i) + \beta_3(Cards_i) + \beta_4(Age_i) + \beta_5(Student_i) + \varepsilon_i \quad (1)$$

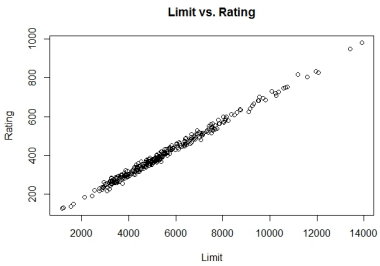
$\varepsilon_i \sim N(0, \sigma^2)$
ε_i = error in the i^{th} person
β = coefficient corresponding to observed variables

This model is developed to predict the i^{th} person's balance, given income, limit, number of credit cards, age, and student status (as an indicator variable). These variables were chosen as the best model to predict the most likely credit balance, given the other variables. The model can be used by collecting these five pieces of information from future clients (assuming this is possible), and plugging them into the model. This would give a predicted credit balance of the future client, allowing the credit card company to assess whether they would be a profitable customer. We will discuss the precision of the model in the explanation of the model cross validation.

2.2 Methods

In deciding which variables to include, and if they require transformations, I worked through a few different processes. The response, credit balance, has the ability to have a negative balance. If a client returns an item that has been paid for with credit, and has already paid the bill, the credit card company can owe the client money. For this reason I do not transform the credit balance to the log scale.

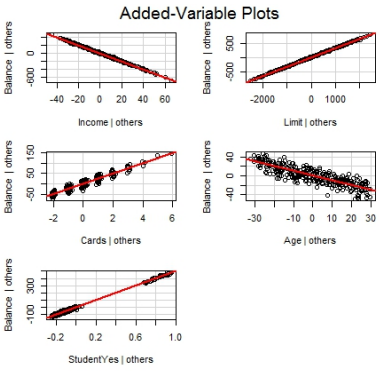
As for explanatory variables, I looked at the relationships of these variables with credit balance, and with each other. After finding the variance inflation factors with all available variables in the model, I have chosen to include credit limit instead of rating, as both had high variance inflation factors, and limit produces a higher R^2 value and a lower mean squared error (MSE). The variables have a correlation of 0.996 (see plot), and will inflate the variance in the model if both are left in. This change improves the residual errors, as well as bolsters predictions.



Other possible explanatory variables, such as the interaction between a clients student status and their income proved to have no significant effect in the model, and were therefore excluded. All other variables included in the model were chosen through comparing AIC, BIC, Mallow's Cp, and adjusted R^2 through a best subset selection process. Each information criteria gave slightly different results. Since my main goal is prediction power from the model, I chose the fewest variables that would still produce accurate predictions with small prediction intervals. In my opinion, this provides a crucial combination of parsimony and prediction ability, both being necessary components of this problem.

3 Assumptions

With any linear regression, certain assumptions must be met. The first assumption is linearity between the explanatory variables and the response, in this case it is our client characteristics with credit balance. In exploring the model, the relationship between balance and the included variable seem to be linear (see plot). This added-variable plot shows the linearity after adding sequential variables to the model. It can be seen that there maintains a linear relationship as each variable is added.

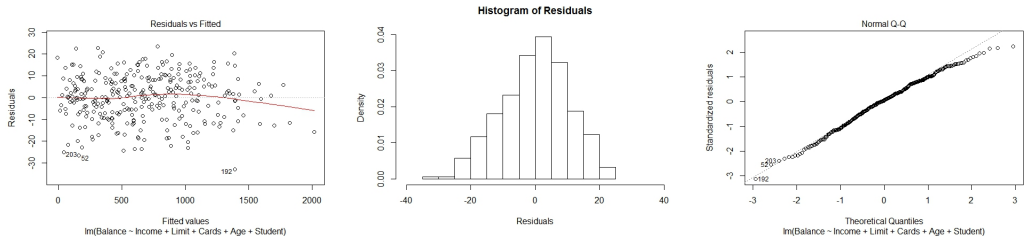


The second assumption, independence between each client's credit balance, really comes from the intuition of the problem. Since each credit card is completely separate from every other, and the balance of one client doesn't directly affect the balance of another client, we can assume independence. To add more validity to this assumption, I will graphically illustrate this with a fitted values vs. residual plot (see plot below). Because there are no clear patterns or shapes we verify the assumed independence.

The errors produced from the model must follow a normal distribution for the third assumption to be met. This can be checked by creating a histogram of the residuals (see plot below). There is a fairly normal shape, allowing us to assume that the errors are normal. This argument is strengthened by the normal q-q plot (see plot below). This plots the residuals against the theoretical

quantiles. We could say that the tails are a little bit different than the theoretical normal, but overall the normality of the errors is not a bad assumption to make.

The last assumption, equal variance, requires that there be uniform spread across all residuals. We can use the fitted values vs. residuals plot to see any heteroskedastic patterns. Since there is even spread among the residuals, we can assume that there is equal variance in our model.



After the four explicit assumptions made above, there are some implicit assumptions that ensure the model's validity. I checked for outliers and influential points, but none were obvious enough to require changes in the data, or the model. Overall, the data is fairly clean, and reasonable when fit to the model.

4 Performance Evaluation & Model Interpretation

4.1 Results

Now that our assumptions are met, we can use the model to make inference, and predictions. We can assume that our estimators can be used to make observations on credit card balance of all individuals for this credit card company. Before stepping further into the model, we will evaluate the model's overall performance. This model has an R^2 of .9993, suggesting that the model explains 99.93% of the variability in credit balance. The model carries much of the information needed to know about credit balance. Each estimate included below were statistically significant, and add value to the model's ability to explain credit balance. The model's MSE is 109.65, much lower than the MSE produced from other models.

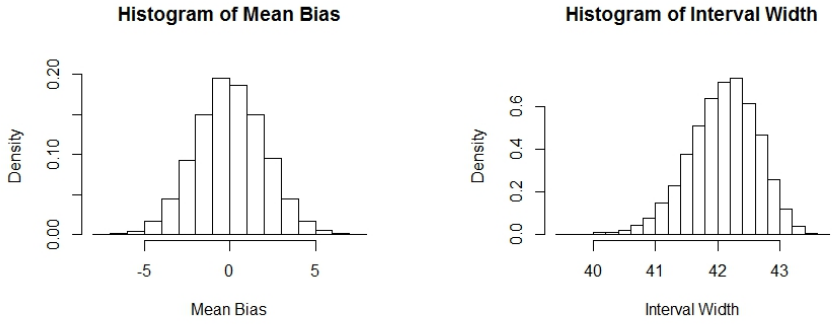
Coefficient	Estimate	Lower 2.5%	Upper 97.5%
β_0 (Intercept)	-701.8	-707.8	-695.78
β_1 (Income)	-9.994	-10.05	-9.935
β_2 (Limit)	0.3264	0.3254	0.3275
β_3 (Cards)	24.81	23.98	25.64
β_4 (Age)	-1.003	-1.073	-0.933
β_5 (Student)	500.6	496.94	504.23

For the sake of making the model interpretable, I will include interpretation for a couple of coefficients. In the chart of coefficients shown above, there are estimates for the coefficients β_i. The estimate for intercept does not carry much meaning, as it would say that the credit balance for someone who is zero years old is -$701.80. This makes the estimate somewhat useless to us. However, other coefficients provide us valuable information about credit balance. For example, the coefficient for income can be interpreted to mean: as a client's income increases by $1,000/month, the clients monthly credit balance will go up by -$9.99 (or go down by $9.99) on average, with 95% confidence that the true value of the coefficient is between -$10.05 and -$9.94. And for a categorical variable, the interpretation would be as follows: as a client becomes a full-time student from not being a full-time student, the clients credit balance will go up by $500.60 on average, with 95% confidence that the true value of the coefficient is between $496.94 and $504.23. These interpretations can be extended to the other coefficients in a similar fashion.

4.2 Prediction Evaluation

We are also interested in how this model can predict the credit balance for future clients. To test this, I ran a cross-validation study to test how well the model predicted a known portion of the data. After taking out a random 10% sample of the data, I ran the model and used the generated estimates to predict values for the missing 10%. As this process was repeated 10,000 times, I

found that the mean bias of each 10% sample was centered around zero, and was between -5 and 5 (see histogram). The prediction intervals from these predictions cover the actual value 95.27% of the time. The prediction values also have an average width of 42.09 (see histogram). In more applicable terms, the predictions are generally within $20 of the actual credit balance. This model gives very accurate predictions of individual credit balances, and has a tight spread of variation, which allows the predictions to be very close to the actual value.



5 Conclusion

With the available variables, this model accomplishes the goals of the study. The model explains the relationship between income, limit, number of cards, age, and student status and the balance of a client. Not only does the model explain the relationship, but it also exploits the relationship by having the ability to predict credit balance with these variables. The model does extremely well at producing accurate estimates of individual credit balance.

The model works extremely well, given the variables that are given in this dataset. One thing that may be an issue when applying this model to future potential clients is that the credit card company may not know the credit limit of the applicant. Since rating and limit are so closely related, the credit card company could look into using rating in their model instead. This does increase the variability of the predictions, but it may be more useful for practical applications. The other variables seem to be information credit card companies can obtain fairly easily through credit rating companies, though the number of credit cards an individual has may not come in that report.

Something that can improve this study is to find the optimal balance for interest rates to be most profitable to the company without putting the client at risk for default on the credit loan. Another piece of this puzzle that could be solved is optimizing the card to the type of user, which would require segmenting the data and understanding what demographics buy what, and how can a credit card get them to the optimal amount of credit balance, making it profitable for the credit card company. This study has many avenues to discover, but the understanding of what drives credit balance has been discovered, as well as an effective method for predicting balance.