STAT 536 Case Study 1: Credit dataset

Jeremy Meyer

May 5, 2019

Abstract

Credit card companies are interested in predicting credit card balance to both minimize risks associated with issuing credit and maximize profit. This analysis will consider a multiple regresson model with credit card balance as the response. Assumptions for the linear model will be checked and problems in the data will be addressed. The interaction effect of student status and income will also be examined. Model performance will then be assessed using cross validation and then we will explore specific research questions such as if people get more responsible as they age. Policies for increasing the credit limit with increasing income will be addressed.

1 The data

Credit card companies are interested in making predictions of customer credit balances before they issue the cards. This is due to some of the risks associated with issuing credit. Card lenders make most of their money on interest, so low balance members pose no risk, but yield little profit. However, extremely high balances pose a huge risk because the card holder may default on their balance, resulting in a huge loss of money. Customers that have moderately high balances are ideal. This analysis will primarily be concerned with predicting credit card balance and making inferences on factors that contribute to higher balances.

The data has 294 individuals with 10 variables each. Some examples of covariates are income, number of current cards, credit limit, credit rating, years of education, and age. Plots of each quantitative variable with the axis names on the diagonals are in Figure 1 on the next page. One problem with the data is that credit limit and rating are highly correlated with each other. One of these two variables may have to be eliminated because it carries some redundant information.

The categorical variables are displayed in Figure 2 on the next page. Balance differences between the 37 students and 257 non-students appear significant because the median student has a higher balance than a 75th percentile non-student. Additionally, since graduation from school may yield a substantial increase in income, an interaction effect between student and income will be considered. Ethnicity, gender, and marital status all appear to have no effect.

1.1 Goals

The goals of the analysis are to address the following research questions:

1. Are the collected variables able to adequately predict a person's credit balance?
2. Do people generally get "more responsible" (in terms of lower balances) with money as they age?
3. Current policy suggests that limit should increase by 10% of an income increase. Under this policy what is the expected difference in credit balance when a person's income goes up by 10,000?

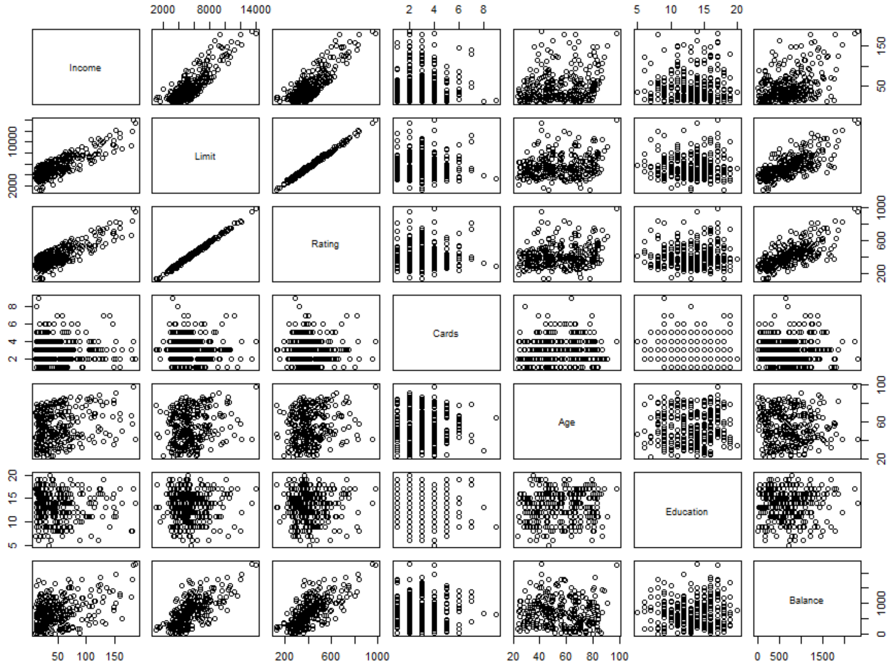


Figure 1: Pairs plots of each numerical variable

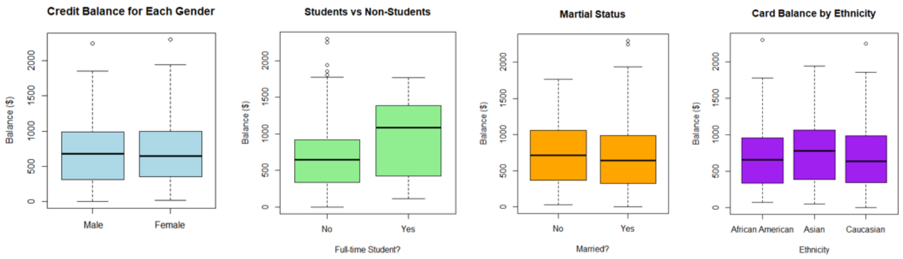


Figure 2: Boxplots of each categorical variable

4. What should the company's policy be for increasing credit limit for increases in income?

Finding characteristics that yield higher balances can give credit card lenders variables to pay attention to and reveal target audiences they can focus on. Predicting the future balance from the cardholder can help credit card companies with some logistical problems like determining how much credit they can afford to lend out.

2 The Model

A multiple linear regression model with balance as the response will be used to analyze the data. A regression model will be useful because of its ability to both predict and make inference on the variable effects. Thus, this can be used to predict credit balance, determine if age is a significant variable balance, and determine policies for optimal credit balances. To begin, first consider the full model (Equation 1) with all variables and an interaction term for student and income:

$$\begin{aligned} Balance =& \beta_0 + \beta_1(Income) + \beta_2(Limit) + \beta_3(Rating) + \beta_4(Cards) + \beta_5(Age) + \\ & \beta_6(Education) + \beta_7(Female) + \beta_8(Student) + \beta_9(Married) + \beta_{10}(Asian) + \\ & \beta_{11}(Caucasian) + \beta_{12}(Income)(Student) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \end{aligned} \tag{1}$$

Each β_i corresponds to the expected increase in balance as its corresponding explanatory variable increases by 1 unit. The categorical variables are taken care of by being turned into binary outcomes. For example, if the person is female, the value for $Female = 1$, and if male, $Female = 0$. If the person is African American, then both $Caucasian = 0$ and $Asian = 0$. The ϵ term is random noise added to the expected balance given all the predictors. It is assumed to be normally distributed with zero mean and fixed variance.

2.1 Checking Assumptions

There are several assumptions that need to be checked for multiple regression. First, the relationship between the predictors and response should be linear. This can be validated using added-variable (AV) plots, which considers how much one predictor affects the response after taking into account the effects of all other predictors. Only quantitative variables were plotted because it does not make sense to test linearity for binary variables. From the AV plots shown below, the relationships look fairly linear.

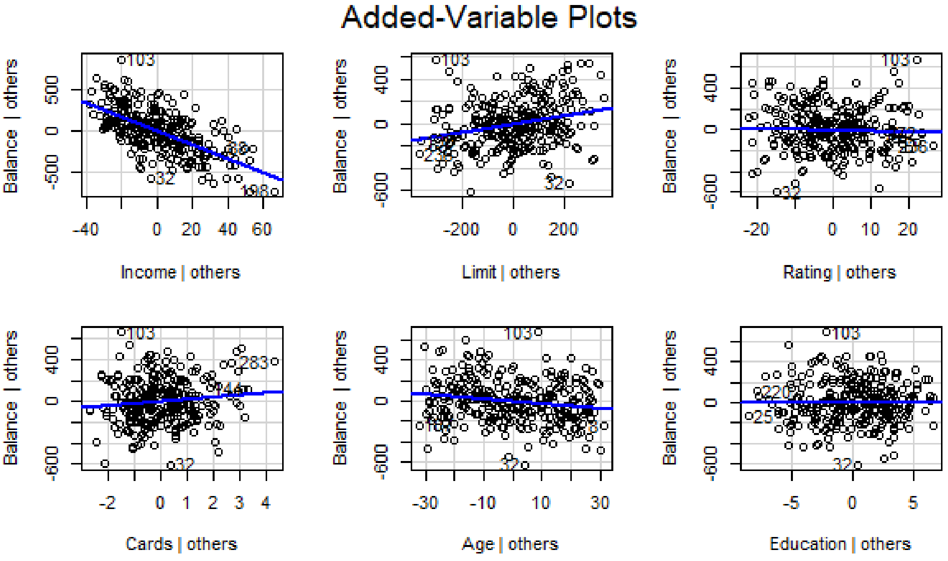


Figure 3: Added Variable plots

Independence between observations is another assumption in the model. This could be violated if the data was not taken from a random sample. An example is that it contained many married couples with similar characteristics. Sometimes structural patterns can be seen by plotting the fitted values against the residuals. The red dashed lines indicate $\pm 2\hat{\sigma}$ from zero. From Figure 4, no clear pattern can be seen, thus the assumption will hold.

Equal variance is another assumption that can be seen from the residuals and fitted values plot. If the values are funneling out or in, the variance is not constant. However, from figure 4, the variance looks stable; the values remain mostly contained in the $\pm 2\hat{\sigma}$ lines.

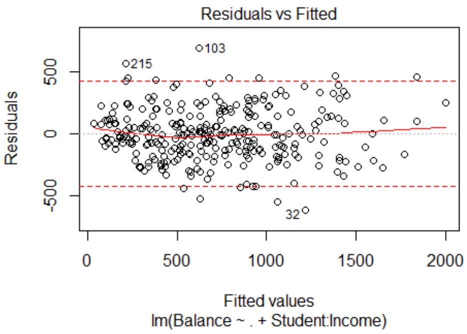


Figure 4: Residuals vs fitted values plot

# Predictors	1	2	3	4	**5**	6	7
Cp Value	362.7	207.38	12.41	4.605	**2.867**	4.09	4.69

The last assumption is normality. These can be checked using a QQ plot or a histogram. Since a lot of the sample and theoretical quantiles match up and the histogram looks very normal, the normality assumption is satisfied.

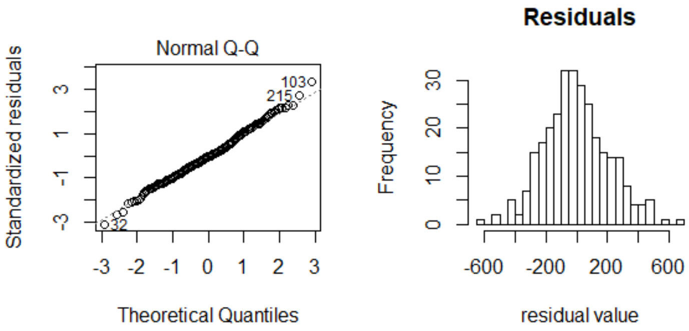


Figure 5: QQ plot and residual histogram

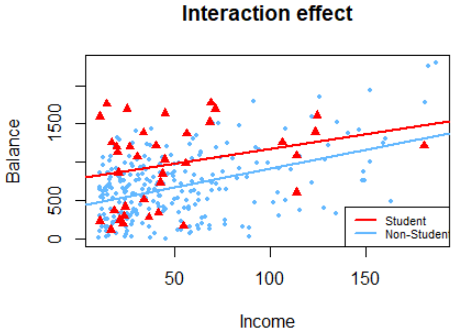
2.2 Variable Selection

One problem mentioned earlier was that limit and rating were highly correlated: in fact, their VIFs are as high as 180. To eliminate unimportant variables, best subset selection will be used because it's computationally feasible for the data. The C_p criterion will be used, which is the same as AIC under the gaussian linear model. This was used so the model will retain more variables that can be available for tests and better predictions. The results are displayed in a table on the following page.

A model with 5 predictors had the lowest Cp. Its covariates were Income, Limit, Cards, Age, and Student. These will be used in the final model. Using the best subset model, 3 main outliers were found, but taking them out had negligible changes on the significance and estimates of the betas. Thus, the outliers could be safely ignored.

2.2.1 Student and Income Interaction

Since students can get a substantial increase in income after graduation, an interaction effect will be considered. However, the interaction was not even selected from the best subset algorithm. Additionally, if it is thrown into the best subset model, it has a p-value of 0.876. We would not conclude this effect is significant. This can be seen by the similar slopes between students and non-students on balance in the graphic below.



2.3 Model Performance

After checking assumptions and performing variable selection the final model is as follows:

$$\begin{aligned} Balance =& \beta_0 + \beta_1(Income) + \beta_2(Limit) + \beta_3(Cards) + \beta_4(Age) + \beta_5(Student) + \epsilon \\ & \epsilon \sim N(0, \sigma^2) \end{aligned} \tag{2}$$

We will now check model performance by cross validation and address the first research question. We will randomly select 70% of the data to train a model using the 5 covariates selected earlier. To check model fit, MSE and bias metrics will be computed for the remaining 30% of the data. This was repeated for 1000 iterations, so the average for all iterations will be displayed. To check predictions, we will compute prediction interval coverage and average prediction interval width.

The average bias was 0.644, which means our estimates are unbiased. The average MSE over 1000 iterations was 46300. If we compare this MSE to the variation from just the mean of balance (or $\sum(y - \bar{y})^2/n$), the model explains about 77% more variation than just the credit balance mean. The 95% prediction interval coverage was about 94.4%, which is about as expected. The average prediction interval width was 841.9. Relative to the range of the balance data, the average interval covers about 36% of possible values. The model fits the data well, and so the 5 collected variables are able to successfully predict a person's balance. However, the variances are a little high for the predictions.

3 Results

Listed in Figure 6 are the beta estimates for the final model. Each beta estimate corresponds to the effect on balance for a 1 unit increase in its corresponding predictor. For example, since income is in thousands, a one thousand increase in income alone corresponds to an expected decrease in $8.51 in credit balance. The fact that the person is a student will on average,

Figure 6: Final model estimates with 95% confidence intervals:

	Estimate	2.5%	97.5%	p-value
(Intercept)	-547.747	-671.868	-423.626	2.87e-16 ***
Income	-8.505	-9.688	-7.322	<2e-16 ***
Limit	0.305	0.283	0.327	<2e-16 ***
Cards	16.592	-0.207	33.394	0.0529 .
Age	-2.336	-3.771	-0.902	0.0015 **
Student	531.829	457.099	606.560	<2e-16 ***

increase credit balance by $531.83. All variables but cards have a significant or non-zero effect on credit balance.

The next research question we will address is if people tend to get more ”responsible” as they age. That is, does a higher age tend to result in a lower credit balance? Based on our model, we would conclude that yes, Age has a negative effect on credit balance because the p-value is significant. However, it is worth noting the expected difference is slight, being only about -$2.34 per year. The effect can be seen on the age AV plot back on Figure 3.

The current policy states that credit limits should increase by 10% of any income increase. We are interested in answering what the expected balance increase is with a $10,000 increase in income. In the model, we want to test what happens when income goes up by 10 and limit goes up by 1,000 simultaneously. The result was an expected increase of $220.38 with a 95% confidence interval of [206.51, 234.26]. Adjusting limit with an income increase has a considerable effect on credit balance. If this is not made with an income increase, the expected balance is actually lower (-$8.51).

To determine an ideal percent of income increases to increase credit limits, consider the table below that has different percent rates corresponding to an income increases of $10,000.

Balance increases for a $10,000 increase in income

Policy % rate	2.5%	Estimate	97.5%
5	61.09	67.66	74.24
10	206.51	220.38	234.26
15	348.86	373.10	397.35
20	490.77	525.83	560.88
30	774.26	831.27	888.28

Based on these results, 10% is a good value for a $10,000 increase in income because it more than triples the expected balance increase from just a 5% rate. If a boost is desired to customers with low balances, a 15% increase may be feasible if companies are willing to risk a potential $373.1 expected increase in credit balances. A 10% rate seems feasible for most cases. Companies could also consider the customer’s current balance and see if a higher rate will boost balances. Also, since a higher increase in income will scale the balance increases, the rate may need to be decreased if a substantial increase in income occurs.

4 Conclusions

The main variables that contributed to credit balance were customer income, existing credit limit, number of owned cards, customer age, and if they are a student. These variables made fairly accurate predictions, but the uncertainty associated with their predictions was high. This means that the numeric interpretation of the effects can be somewhat volatile depending on the data it’s trained on. Based on the data, we found a slight negative relationship between age and credit card balance, suggesting that older people are more responsible with their credit cards.

We also found that the current policy of increasing limit by 10% of any income increases had an expected increase of $220.38. There was no clear reason to change this, unless companies wanted a larger boost in balances.

One potential concern in the data is how representative it is and if it is able to be applied to multiple regions. Maybe the data only reflects current economic or cultural trends. One question that may be of further interest is what constitutes a risky ”high” balance. Finding customers that may potentially default on their loans from existing balance data can save card companies a lot of trouble. If companies can know which customers are likely to default on their loans in the next few years, companies can know when to curtail credit limit increases or be more aware of specific risks they are taking.