Title: Interest rate is associated with loan length for similar FICO ranges.

Introduction:

An interest rate is the rate at which interest, what can be thought of as the "fee" for a loan, is paid by the borrower to the loaner. Interest rates depend on a number of different factors like the current economic situation, political motivations, or risk [1]. One way to determine the risk on a personal loan to an individual is by the credit worthiness. "The best-known and most widely" used measure of credit worthiness, and the one used in this project, is the FICO score. FICO scores are based on a number of different factors including the way a person uses credit, the length of a person's credit history, and the individual's history of paying bills [2].

Having a good understanding of how interest rates are determined is crucial to developing good financial practices and responsibility. In this project we performed exploratory analysis and used standard linear regression methods to determine what factors affect interest rates for individuals with similar FICO scores. We were able to find a significant relationship between the length of a loan and interest rates and suggest that longer loans are associated with lower interest rates.

Methods

Data Collection

For our analysis, we used data collected by the Lending Club [3] and provided to us by our course instructor (information about date downloaded and exact source are therefore unavailable) [4]. The data was downloaded using the R programming language [5].

Exploratory Analysis

We preformed exploratory analysis on the data by creating plots and reading tables from the data. We changed the type of data when necessary (mainly transitioning to numeric values rather than factor) and made sure the data contained values that were consistent with background knowledge. This part of the analysis was also crucial in determining what type of linear model to apply to the data.

Statistical Modeling

To determine the relationship between loan length and interest rate we used a standard multivariate linear regression model. This model was chosen based on the exploratory analysis performed beforehand. The coefficient, p-values, errors, and confidence intervals were calculated using the R programming language's built in "lm" and "confint" functions.

Reproducibility

All of the procedures and charts were logged in an R markdown file that ensures a reproducible code with extensive commenting to help guide the reader. The data we used was only available to us through the course website but can be found in the references section.

Results

The data provided by The Lending Club contains 14 different variables for each individual loan and consisted of 2,500 observations. Due to a lack of a strong relationship in the other variables during exploratory analysis, only the interest rate (interest.rate), FICO scores (fico.ranges), and the length of the loan (loan.length) were used. It should be noted, however, that not finding relationships during exploratory analysis does not mean there weren't any others (more on this in the conclusion).

Of the selected variables, there were no missing values and all values were within expected ranges. 78% of all loans were 36 months long while the remaining 22% were 60 months long. There was no apparent relationship between loan length and FICO score, which again agrees with our background knowledge and also means loan length is not a confounder. We then created a linear model in an attempt to explain all the patters of the data and developed the following model:

$$interest.rate = b_0 + b_1*fico.range + h*loan.length + g*loan.length*fico.range + e$$

where b_0 represents the intercept term, b_1 represents the change in interest rate per one unit of a FICO score across the same loan length, the h*loan.length term represents the two levels of the corresponding factor variable, the g*loan.length*fico.range accounts for any interactions the two variables have, and e, or error term, represents any unmodeled random variation in interest rates, most of which appears to have been removed from our model.

Our model led us to observe a statistically significant ($P < .001$) relationship between interest rate and FICO score. For each unit increase in FICO score there was a -.085 change in interest rate (95% confidence interval: -0.087, -0.082). Additionally, the difference between interest rates for a 30-month loan and a 60-month loan as FICO scores increased by one unit was -0.0167 ($P < .001$, 95% confidence interval: -.023, -.011). In other words, when FICO scores increased the interest rate for a 60-month loan decreased at a faster rate than interest rates for 30-month loans.

Conclusion

We concluded that interest rate was clearly related to FICO score and that loan length had a direct impact on interest when FICO range was taken into account. This result is important when considering personal finance. We essentially learned that if

you have a higher FICO score, the length of your loan with have a lower impact on your interest rate.

While this conclusion is may prove useful to some, there are a variety of other variables that may not have been accounted for. During our exploratory analysis we failed to show that any other given variables had a significant influence on interest rate given a certain FICO score. But, as stated earlier, this does not imply such relationships do not exist. Perhaps our analysis simply failed to properly demonstrate different relationships. Or perhaps the data did not include all significant variables.

Additionally, there was not too much background information on the data used for the analysis. Information such as the time and date the data was collected and the exact methods of data collection were all unavailable. It should also be noted that I only vaguely know what I'm doing because I just learned this material less than a week ago. Cheers!

References

1. Wikipedia, "Interest Rates." http://en.wikipedia.org/wiki/Interest_rate . Accessed 2/16/13.
2. Wikipedia, "Credit Score." http://en.wikipedia.org/wiki/Credit_score_in_the_United_States. Accessed 2/16/13.
3. Lending Club Website, http://www.lendingclub.com/
4. Coursera, Data Analysis Course, (URL unavailable due to restricted access).
5. R Core Team (2012). "R: A language and environment for statistical computing."   http://www.R-project.org