Karen Yang

Title: The “Usual Suspects” Affecting the Interest Rate?

Introduction:

There are many factors that impact the interest rate for loan-seeking individuals. This study focuses on four commonly analyzed factors: FICO score, monthly income, debt-to-income ratio, and length of employment. Do these still matter as explanatory factors in a non-traditional setting such as peer-to-peer lending? This study uses data from one lender, namely the Lending Club, which is a peer-to-peer lender and tells a story about whether “the usual suspects” have a role in determining the interest rate in a non-traditional lending setting [1].

The purpose of this analysis is to identify and quantify associations between the interest rate of the loan and the four other variables mentioned. In particular, I consider whether any of these variables have an important association with interest rate after taking into account the applicant's FICO score, a very important score because it is a measure of credit risk [2].

Understanding the relationship between the interest rate and FICO score along with other variables serves a very important and practical purpose. It can help shed light on the factors that affect loan rates in a non-traditional, peer-to-peer lending setting. The Lending Club’s data show that the loan terms are generally short term—either 36 or 60 months. The ideal interest rate is the lowest rate that one can obtain in order to minimize the cost of borrowing money.

I conducted an initial bivariate regression analysis to determine if there was a significant association between the interest rate and the FICO score range. I further used a multiple regression analysis, incorporating three additional variables, namely monthly income, debt-to-income ratio, and length of employment. Much like the bivariate analysis, the multivariate analysis showed the FICO score as statistically significant.

Methods:

Data Collection

For the analysis, I used data from the Lending Club that consists of 2,500 peer-to-peer loans issued through this lender’s website [1]. The data were downloaded from the following website on February 8, 2013 using the R programming language [3]: https://spark-public.s3.amazonaws.com/dataanalysis/loansData.rda. To see the raw data directly, the data can be viewed at this website: https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv. The codebook, which explains the variables and their measurements, can be accessed at here: https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf. Unfortunately, loan issuance dates are not included in the data obtained, though it could possibly be an important factor. It is advised to keep this in mind as a possible confounding factor that is missing from this analysis.

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. We identified transformations to perform on the raw data on the basis of plots and knowledge of the scale of measured variables. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, and (3) determine the terms used in the regression model relating the interest rate to the other identified variables.

With 2,500 observations and 14 variables in a single sample dataset, there were 7 missing variables found in 2 rows that were omitted. Starting with the outcome variable, namely the interest rate, the mean interest rate for this dataset was approximately 13% with minimum set at 5% and the maximum set at 25%.

In further describing the data, the mean loan amount requested and funded was $10,000, which reflects a relatively small loan. Interestingly, this amount is close to the mean amount in revolving credit card balance, that of $10, 962. The debt-to-income ratio showed a minimum and a maximum of 0% to 35% with a normal distribution. In terms of inquiries into credit history, 1907 out of 2500 individuals had 1 or less inquiry. Based on this single fact, this variable does not seem to have much variation to make it worthwhile to include in the analysis for either statistical or substantive meaning.

As for the length of employment, 653 had 10 or more years of employment while it was less than 250, each respectively, for one to nine years of employment as well as less than a year's employment. There were 77 observations for the category of "n/a", which was not specified in the codebook. Hence, this category's meaning is unknown.

In cleaning the data, the "%" was removed from the interest rate variable column as was "months" from the term of the loan variable, denoted as loan length. Other small changes were made such as making the data numeric for the interest rate, the debt-to-income ratio, and the FICO score. For monthly income as measured in US dollars, a log transformation and an addition of 1 was made to the column of income data because the distribution was found to be right-skewed. This is a standard practice to help normalize data for purposes of computation and analysis.

For preliminary data analysis, plots and bivariate regressions were made to assess the relationship between each of the 13 explanatory or predictor variables with the outcome or dependent variable, denoted as interest rate. Of all the predictor variables, the FICO range variable had the strongest measure of relationship with the interest rate variable. As Figure 1 shows, the correlation is -0.71, which indicates a strong, negative relationship. Further, the bivariate linear regression model showed statistical significance (P=<2e-16) with the coefficient equal to -4.235e-03, suggesting a small, negative impact on the interest rate. Keep in mind that the FICO score is in a FICO score range between 640-644 to 845-850 and is then categorized by factor levels 1 to 38 to reflect these ranges and is treated as a numeric variable for purposes of computation and analysis.

Statistical Modeling

To relate the interest rate variable with the other variables, I ran a standard multivariate linear regression model. Model selection was performed on the basis of our exploratory analysis with histogram distributions, plots, and bivariate correlations along with some familiarity of the lending process and requirements. Coefficients were estimated with ordinary least squares.

Results:

The Lending Club data used in this analysis contains information on the interest rate offered, measured in percentage as a numeric (INT_RATE), the FICO score range, as measured in factor levels between 1 to 38 that represent FICO ranges of 640-644 to 845-850 (FICO), the monthly income as measured as a log transformation of monthly income in $US with the addition of 1(MI), the debt-to-income ratio as a numeric measure (DI_Ratio), and the length of employment as a factor variable with 12 levels, representing a certain number of years of employment (EMP_LENGTH). As mentioned earlier, I identified a factor level denoted as “n/a” but the codebook did not specify its meaning so the decision was made to keep it in the analysis since it represented 77 observations, even though this factor level is vague and ambiguous. The other data collected for the same 77 observations would still be useful in other analyses.

I used a regression model relating interest rate to four other variables. These four were chosen because they are the “usual suspects” in lending studies that deal with traditional lending institutions. The purpose of this study is to see if these same explanatory factors hold in a peer-to-peer lending scenario, which is a non-traditional setting. Below is the specified model.

$$\text{INT\_RATE} = b_0 + b_1(\text{FICO}) + b_2(\text{MI}) + b_3\,(\text{DI\_Ratio}) + b_4\,(\text{EMP\_LENGTH}) + e$$

Here, we see that b_0 is an intercept term and b_1 represents the change in the interest rate as a percentage with a change of 1 unit of the FICO score range as measured as a factor variable with levels 1 to 38. The coefficient b_2 represents the change in the interest rate with a unit change in the log transformed monthly income variable (MI). Similarly, b_3 is the coefficient that shows the change in the interest rate associated with a 1-unit change of the debt-to-income variable (DI_Ratio), which is a ratio measurement. The last coefficient in the model, b_4, represents the change in interest rate associated with a unit change in length in years of employment as measured by the reference level, which is factor level 1, namely less than a year’s length in employment. Comparison to the reference level is used as “the unit of change”. The error term e represents all sources of unmeasured and unmodeled random variation in interest rate.

As with the bivariate linear regression model that showed the FICO variable to be statistically significant in its effect on the interest rate, I again observe a highly statistically significant (P = <2e-16) association in the multivariate regression model. The coefficient size, however, is very small, that of -0.004291. The p-values for the monthly income variable ((P = <2e-16) and the debt-to-income ratio variable (P = 0.0272) are both statistically significant. The coefficients are 0.01011 and 0.01766, which are both small yet positive values. The changes in interest rates are not affected greatly by a unit change of the log transformation of monthly income or by a unit change of the debt-to-income ratio as measured by a ratio. The length in years of employment is only statistically significant for the third factor level on the employment length variable, which represents 10 or more years of employment (P = 0.0139). The coefficient is 0.005375, indicating a very small effect on the interest rate for moving from the reference level of less than a year's employment to 10 or more years of employment.

Conclusions:

Our analysis suggests that there is a significant, negative association between the interest rate and the FICO score range as a measurement of credit risk, which validates our expectations. This statistically relationship held up in the multivariate analysis in which three additional variables such as the monthly income, debt-to-income ratio, and the length of employment were added to the model. In the multivariate model, the FICO range, monthly income, and the debt-to-income ratio showed statistical significance, though effect sizes were relatively small. However, it was the fourth explanatory variable, namely length of employment years as categorized in factor levels, that only did so for the factor level that had 10 or more years of employment. None of the other remaining factor levels for employment length had statistical significance. Why not?

While our analysis is an interesting first step, it is only based on a limited sample from a single peer-to-peer lender, namely the Lending Club, which is a non-traditional lending outlet. A larger collection of peer-to-peer lending organizations may better shed light on the relationship between interest rate and the other variables in the analysis, particularly employment history as measured in years of employment.

Another perplexing result is that the coefficient sizes of the "usual suspects" were all relatively small. It could be the case that other variables needed to be put in to the model. Thus, there are limitations in this analysis.

This analysis may be of interest to those seeking to better understand factors affecting the interest rate offered by non-traditional lending outlets such as the Lending Club. Perhaps individuals seeking loans through non-traditional avenues such as peer-to-peer lending do so because they do not have an extensive employment history. The data show that of the roughly 2500 observations in this study, only 653 had 10 or more years of employment. We can infer that the majority of individuals seeking loans through the Lending Club have less than 10 years in employment history. This factor may be of interest for a future study to investigate more fully the growing trend in peer-to-peer lending to serve the borrowing needs for those with less than 10 years in employment history. More data from peer-to-peer lending organizations are needed to flesh this out.

References

1. Lending Club Page. URL: https://www.lendingclub.com/home.action. Accessed 2/8/2013.

2. Wikipedia “FICO” Page. URL: http://en.wikipedia.org/wiki/FICO. Accessed 2/17/2013.

3. de Vries, Andrie, and Joris Meys. *R for Dummies.* John Wiley & Sons, 2012.