**Title: Interest Rate Estimation**

**Introduction:**

An Interest rate is the amount charged, expressed as a percentage of principal, by a lender to borrower for the use of assets. The amount of interest rate is depends on many factors such as liquidity preference of lenders, risk of investments, taxes, and even political short-term gains.

Fair Isaac Corporation (FICO) is a type of credit score, which is used to assess an applicant credit risk. It is calculated from different part of data in credit report such as payment history, amount owed, Length of credit history etc.[1,2]

The purpose of this analysis is to identify and quantify associations between the interest rate and the other variables after considering the applicant’s FICO score.

**Methods:**

*Data Collection*

For our analysis we used the data from lending club data set from January 24 2013 to November 14 2013. The data were downloaded from [https://www.lendingclub.com/home.action](https://www.lendingclub.com/home.action" \o "Link: https://www.lendingclub.com/home.action" \t ") on November 16 2013 using the R programming language [3].

*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 for finding other important variables rather than FICO score that affect interest rate.

*Statistical Modeling*

To find the relation between the interest rate and other variables in credit report we performed a stepwise multivariate linear regression method [4]. All possible regression approaches (all includes Fico Score) considered all possible subsets of variables and chose the best model that fit the data. The selection made base on Akalke Information Criterion (AIC).[5]

*Reproducibility*

All analyses performed in this manuscript are reproduced in the R markdown file earthquakesFinal.Rmd [6]. To reproduce the exact results presented in this manuscript the coursera version of the analysis must be used.

**Results:**

The data used in this analysis contains has 13 different properties that may affect the interest rates. Those are amount requested, amount funded by investors, interest rate, loan length, loan purpose, debt to income ratio, state, home ownership, monthly income, FICO score range, number of open credit lines, revolving credit balance, number of inquiries in the last 6 months, employment length.

We identified two rows containing missing data. We collected and all measured variables were observed to be inside the standard ranges. Loan length were either 36 or 60 months and the ratio of populations was about 3 (in the same order). About 50% of population never done any credit inquiry and about 0.2% done more than 8 credit inquiries in the last 6 months.

We first fitted a regression model relating interest rates to FICO scores. The random error of this model was about 2.94. We also fitted a regression model relating interest rates to all variables. The random error of this model was about 2.07.

As a next step we used stepwise linear regression model. Starting from only Fico score as variable, we gradually added to complexity of models. The best model base on AIC score was the following:

***Interest rate ~ FICO score range + Loan length + Amount funded by investors +***

***Number of inquiries in the last 6 months + Number of open credit lines +***

***Loan purpose + Revolving credit balance + Amount requested +***

***Home ownership status***

The AIC and random errors for this model were 3652.32 and 2.07 respectively.

**Conclusions:**

Our analysis suggests that among all those variables in this model FICO score is the most important one. As figure 1 suggests, after FICO score the most important variable is the length of loan. By adding the “length of loan” to the regression model ,which relates interest rates to FICO scores, the value of random error dropped from 2.94 to 2.32.

While as a first step linear model can capture some aspects of relationship between the interest rate and the variables in data set, higher order models can capture more interesting relationship between those. However there is risk of overfitting for those models, which can relieve by testing over a larger data set.

**References**

1. Wikipedia “Interest rate” Page. URL: <http://en.wikipedia.org/wiki/Interest_rate>. Accessed 11/16/2013.

## 2. Investopedia “Definition of 'FICO Score'” Page. URL: <http://www.investopedia.com/terms/f/ficoscore.asp>. Accessed 11/16/2013

3. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: [http://www.R-project.org](http://www.r-project.org)

4. Quick-R “Multiple (Linear) Regression” Page. URL: <http://www.statmethods.net/stats/regression.html>. Accessed 11/16/2013.

5. Wikipedia “Akaike information criterion” Page. URL: <http://en.wikipedia.org/wiki/Akaike_information_criterion>. Accessed 11/16/2013.

6. R Markdown Page. URL: <http://www.rstudio.com/ide/docs/authoring/using_markdown>. Accessed 1/31/2013