I tested all variables against interest rate and came to the conclusion that FICO score, loan length, and employment length to be the more important variables. FICO score had the strongest correlation with interest rate, but loan and employment length both provided a fairly good estimate because there was some correlation between the variables. Note that loan length and employment length are both factor variables. [See later pages for pictures of all of these. The graphs indicate that these variables have some correlation with interest rate, hence their designation as important.]

My R function file contains 7 different functions. The first 5 are simple prediction functions: you pass in a linear model between interest rate and the variable the function is equipped to deal with (its name makes it obvious which one that is) and then a value for said variable, and the function returns the first prediction value of many, not unlike the example provided to us. The first 5 functions handle predictions using FICO score, loan length, state, loan purpose, and employment length.

The sixth function, dual, which now has an incredibly misleading name, was my first attempt to use multiple variables and linear models in the hopes that I could create a better prediction using more data. Dual eventually grew to handle five different variables, but this became cumbersome at best. It created the models in the same manner that the first five functions did. Function calls for dual required passing five different linear models and five different variable values…one for each of the five variables mentioned above (not coincidentally, as this was really just me mashing everything together). The latter part of the function aimed to make a better prediction by averaging all of the (first prediction) values together, but this didn’t seem to work as well as I’d hoped. But the FICO score seemed like the best estimator, so I added it many times to the averaging function, but the results weren’t as impressive as I’d hoped. So I set to work on the seventh function, deciding that there was not enough time left this Tuesday evening, and hoping to have something to show besides a \*good prediction algorithm.

The seventh function, modular, is exactly that. It takes as input all of the data, and then builds a linear model for every variable against interest rate. The second input this function takes is a list, which allows you to be flexible in calling it: you can use as many of or as few of these linear models as you’d like. The list format is as follows:

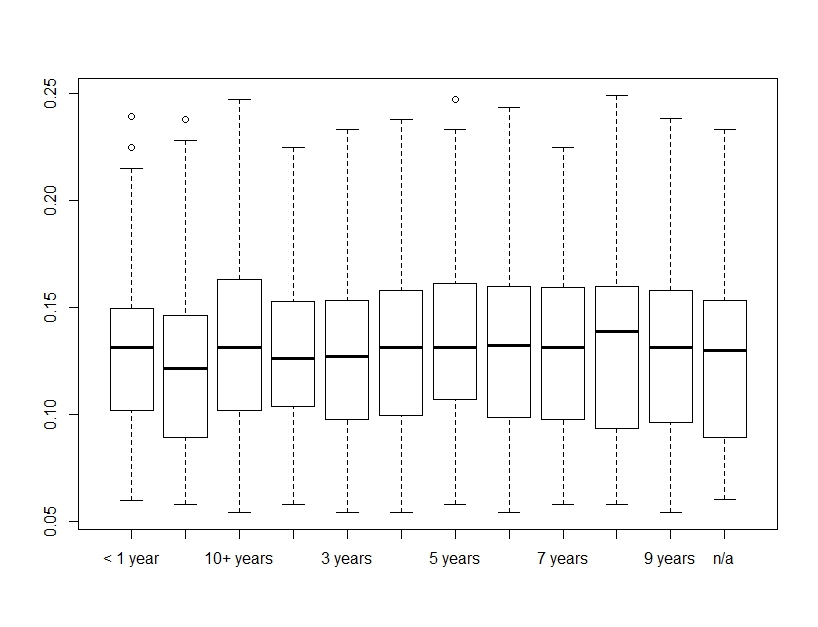
(A variable name from the loan data, exactly as it appears; a value for that variable)

And this set of two repeats. Ex: modular(ld,c("Loan.Length","60 months","FICO.Score",750))

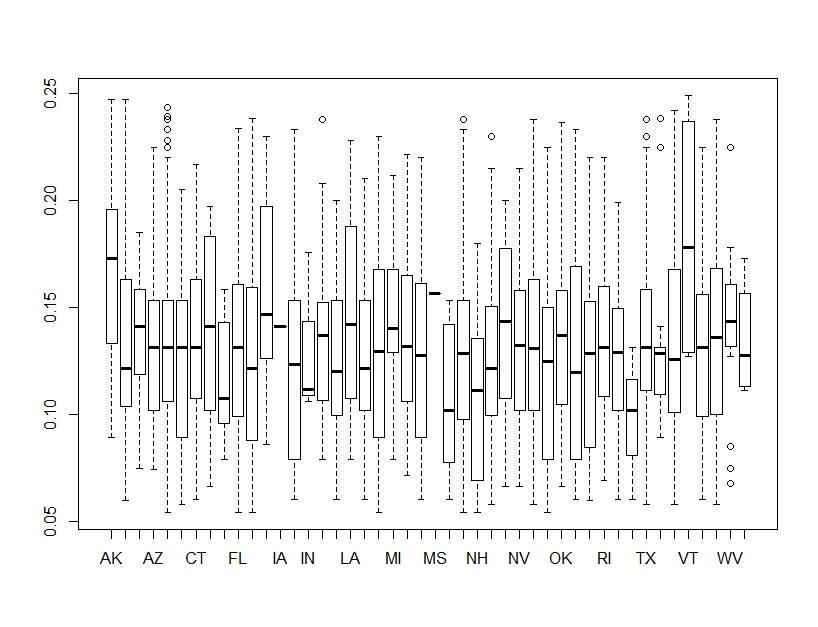
The function takes in any number of variables you’d like, and returns (prints, actually) the prediction for each of them. And in printing, it also labels the prediction with the variable used and the value used. I learned quite a bit about how R handles these data.frames, and frequently encountered a warning where R found many more rows than I intended it to (that description is vague and brief, but if you run into it, you’ll recognize it). Research revealed to me that rewriting the linear model creation as

lm(var1 ~ var2, data) instead of lm(data$var1 ~ data$var2)

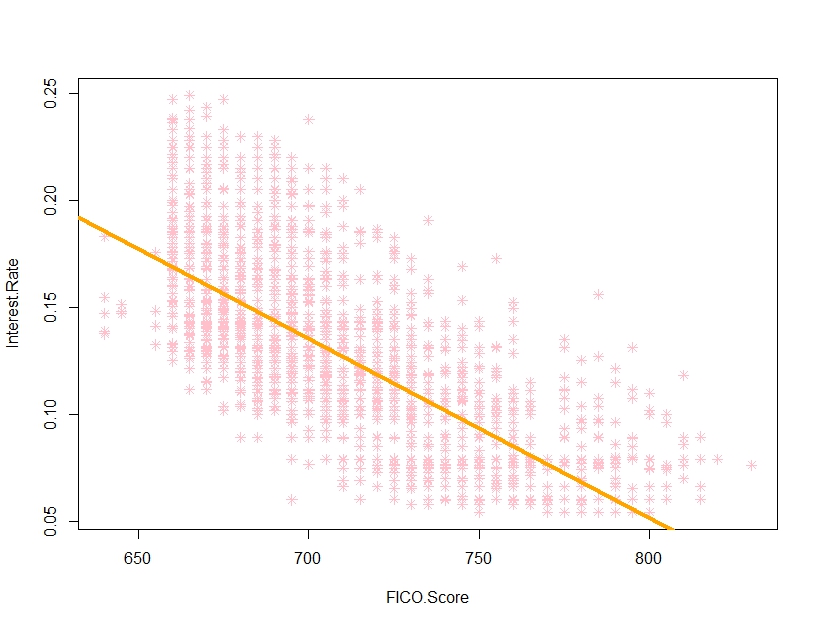
stopped this from happening. I hope you enjoyed reading! It certainly was a fun exercise to make this.



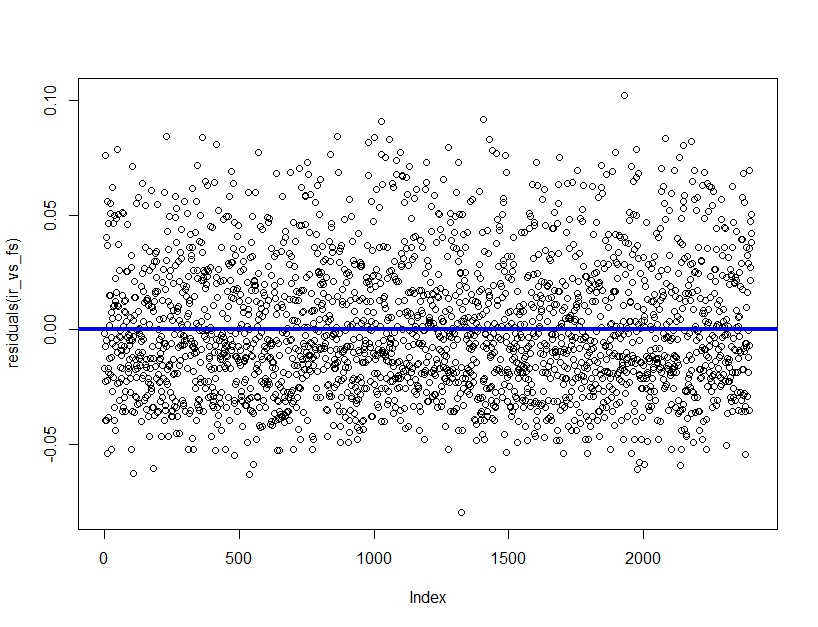
Employment Length vs. Interest Rate



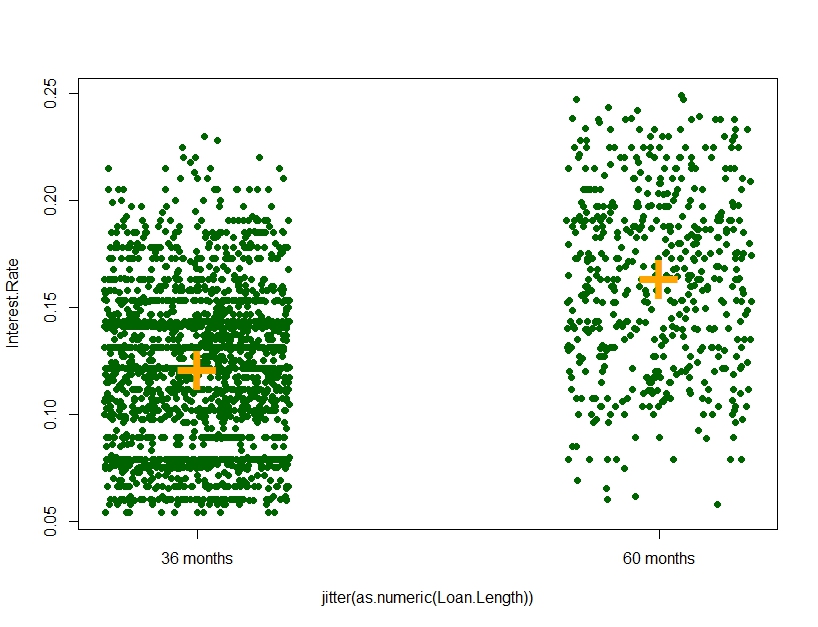
State vs. Interest Rate. Amusing!



FICO Score vs. Interest Rate



Residuals of FICO Score vs. Interest Rate



Loan Length vs. Interest Rate

Residuals:

Min 1Q Median 3Q Max

-0.105517 -0.032082 0.000118 0.025418 0.108418

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.1210825 0.0008685 139.41 <2e-16 \*\*\*

as.factor(Loan.Length) 60 months 0.0423345 0.0018587 22.78 <2e-16 \*\*\*

Residual standard error: 0.03762 on 2398 degrees of freedom

Multiple R-squared: 0.1778, Adjusted R-squared: 0.1775

F-statistic: 518.7 on 1 and 2398 DF, p-value: < 2.2e-16

Tukey multiple comparisons of means

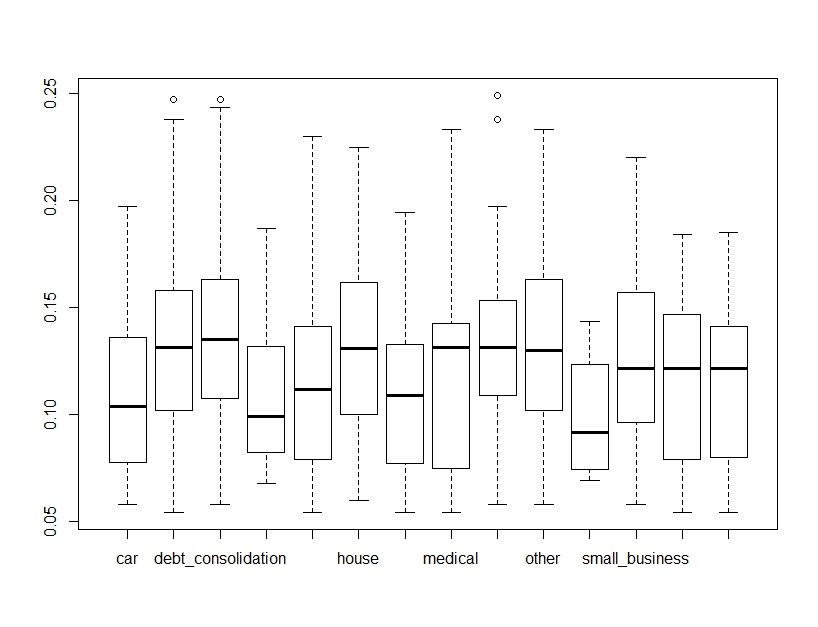
95% family-wise confidence level

Fit: aov(formula = Interest.Rate ~ as.factor(Loan.Length))

$`as.factor(Loan.Length)`

diff lwr upr p adj

60 months-36 months 0.04233452 0.03868961 0.04597944 0



Loan Purpose vs. Interest Rate