Final Project of MA5771 – Applied Generalized Linear Models

Travis Froberg
Analyzing Interest Rates for IBRD Loans Based on Amount of the Loans, Region, and Loan Status

Section 1 Introduction

One function of the World Bank is to provide loans to developing countries. This is ostensibly done to provide these countries with money to improve their economy. The following analysis uses the interest rate of loans from the International Bank for Reconstruction and Development (IBRD) as the response variable. The World Bank website says the following about the IBRD:

"The world's largest development bank, IBRD provides financial products and policy advice to help countries reduce poverty and extend the benefits of sustainable growth to all of their people."

The dataset used in this analysis was obtained from the World Bank website's data repository. Detailed information on how the data was obtained is unavailable. The exploratory variables used in the final model are: the amount of the loan that was disbursed at the time of data collection, the original principal amount of the loan, region of the country who received the loan, and loan status (either cancelled, fully repaid, or active).

The World Bank has been criticized for not living up to its self-proclaimed purpose. There have been accusations that the World Bank lends high-interest loans to vulnerable countries for a variety of reasons. The following analysis examined a sample of loans given out by the World Bank to developing countries (a country must be developing in order to receive funds from the IBRD) to examine whether the World Bank is engaging in predatory lending practices.

In addition, the region of the loan-receiving country was used as an exploratory variable to examine whether the World Bank is engaging in racist lending practices. For example, if it is shown that countries in the region of east and west Africa receive loans with significantly higher interest rates than countries in Europe or the Middle East, then racism can be a hypothesized reason for the difference.

Section 2 Statistical Methods

The software used in this analysis was R Studio 2021.09.0 Build 351. The significance level used was .05, corresponding to a 95% confidence level.

Section 2.1 Exploratory Data Analysis

To get an idea of the relationship between the response and the explanatory variables a scatterplot was created of the interest rate against each numerical explanatory variable. For categorical explanatory variables a boxplot of interest rate against each of them was created.

In addition, a scatter plot and density plot were created for interest rates. This was done to get an idea of which generalized linear model to use.

Section 2.2 Generalized Linear Model

Multiple glms were fitted to the data. These included the gamma and inverse gaussian models because the response variable was positive and continuous. These models did not perform well because the response was highly skewed due to containing a large amount of zeros. The best glm was a tweedie GLM with an index parameter of 1.2769. This tweedie glm was used because it performs well for positive continuous data with many exact zeros.

An F test was used to test the significance of the variable Original Principal Amount and Cancelled Amount.

The overall significance of each coefficient was calculated using a Wald test with t-distribution test statistics.

The assumptions of the final tweedie model were then checked using a plot of the quantile residuals against the fitted values transformed to constant information scale, a plot of the working responses against the linear predictors, a Q-Q plot of the quantile residuals, and a scatter plot of the Cook’s distance.

Section 3 Results and Conclusions

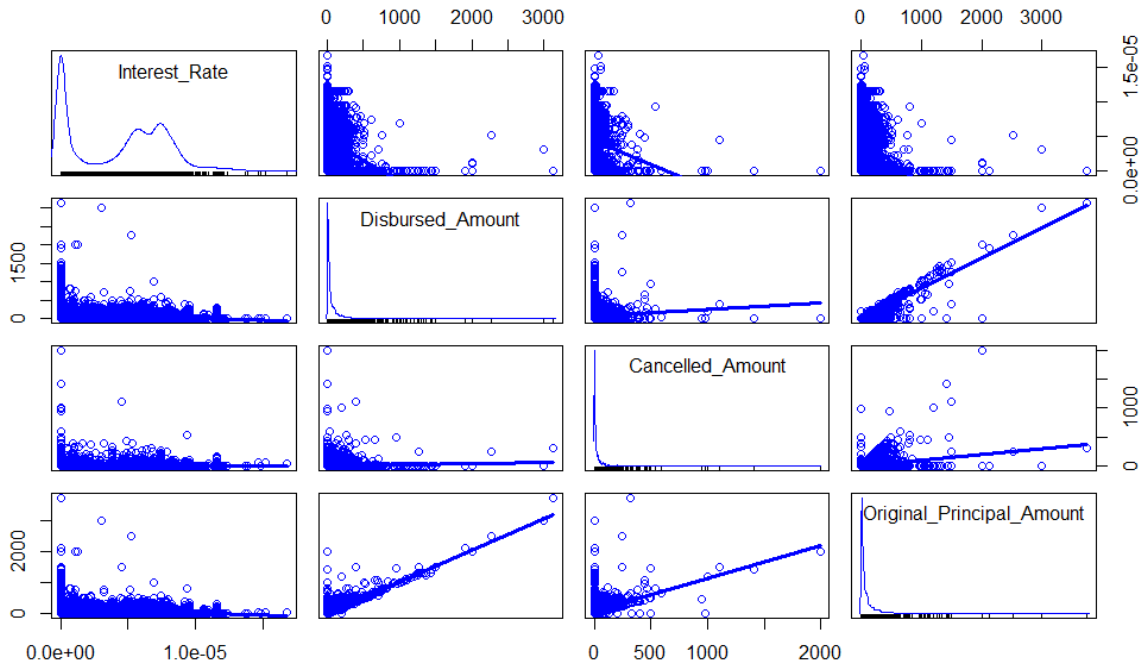

Fig. 1 Scatter plots for all continuous variables involved in model building as well as the response variable (interest rates). Also included are the pdf's for each variable.

As can be seen in Fig. 1, the probability density function for the response variable Interest Rate consists of a bell shape with a positive mean, but is highly skewed to the right because it contains a lot of zeros. A tweedie glm with an index parameter 1.2769 was used as the final model because it performs best on positive continuous data and can handle a lot of zeros.

It can also be seen in Fig. 1 that there are some correlations among explanatory predictors, however, it was not believed that these correlations would affect the response. A few interactions between exploratory variables were tried anyway, but did not improve the model.

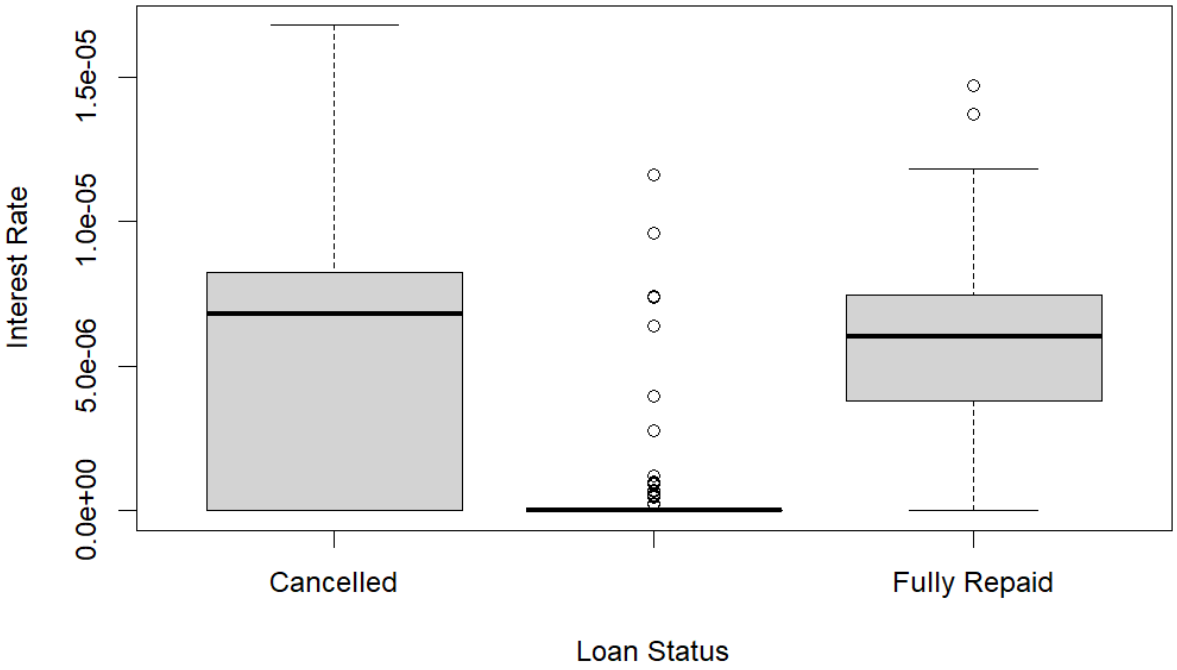

Fig. 2 Boxplot of the response variable (interest rate) against the categorical explanatory variable Loan Status.

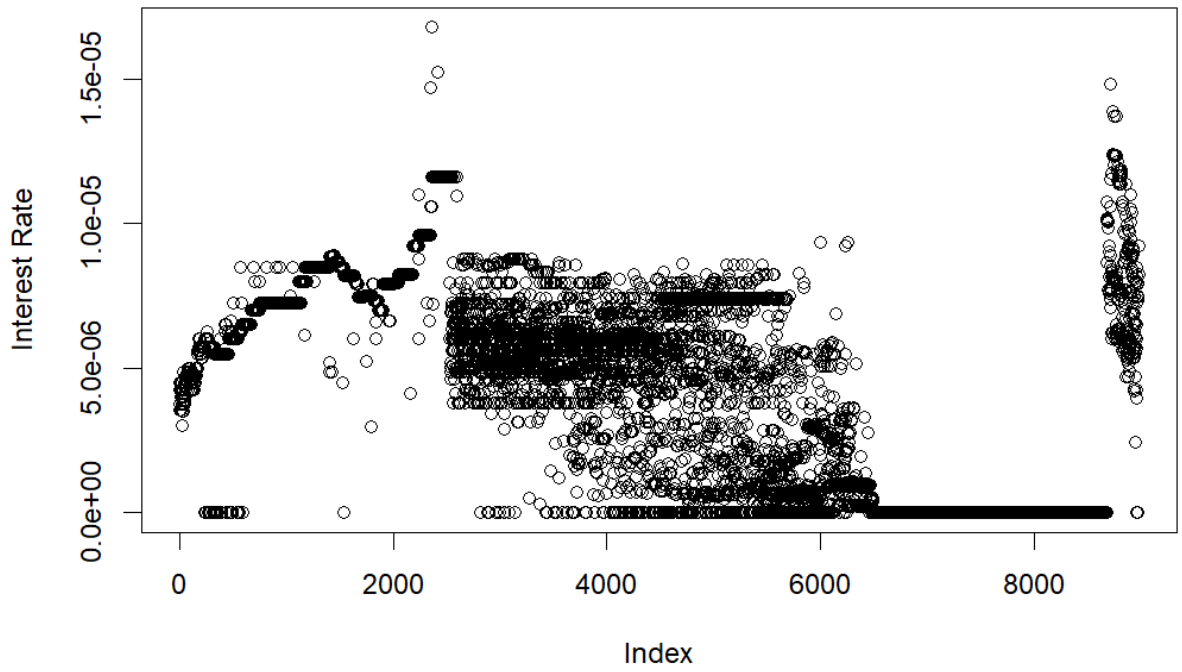

Fig. 3 Scatter plot of the response variable (interest rate)

As can be seen in Fig. 3, the response variable took on some specific values more than others, even though it is continuous. This can be deduced from the horizontal lines in the plot. Poisson and Binomial glms were fit because of this discrete nature of the response variable, but they did not perform well.

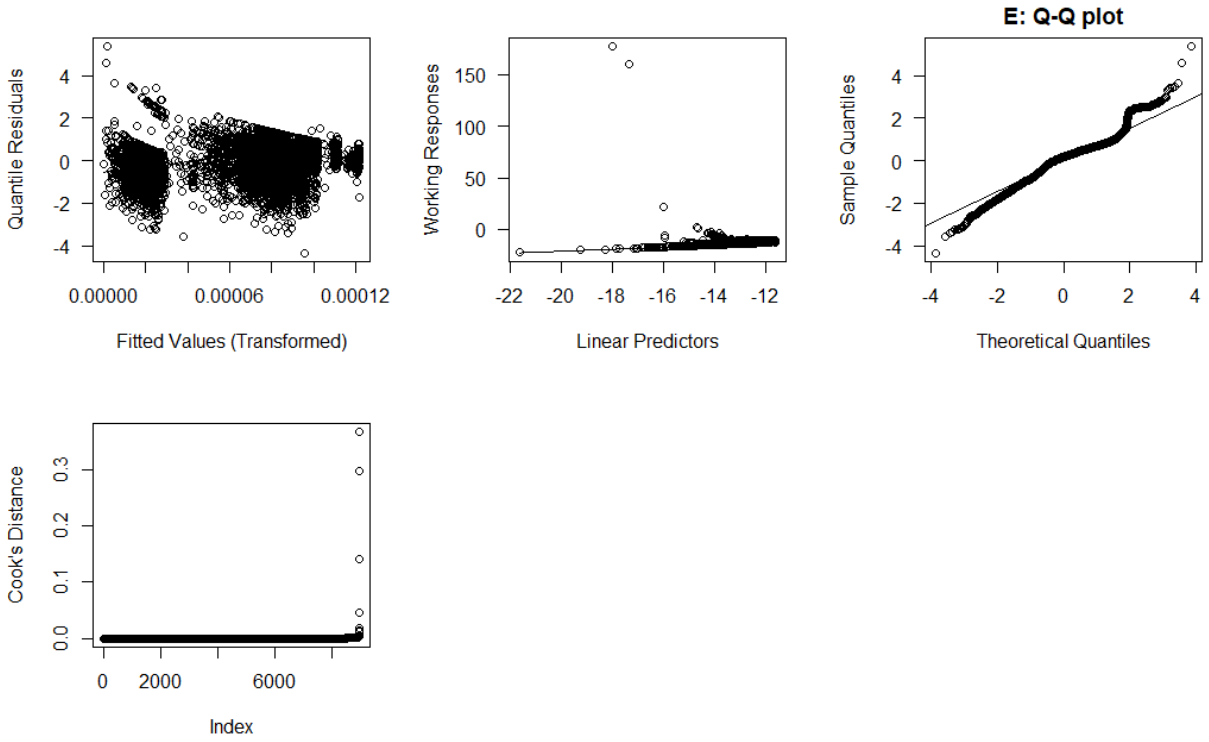

Fig. 4 Diagnostic Plots for Final Model: Top-Left: Quantile Residuals vs. Transformed Fitted Values, Top-Middle: Working Responses vs. Linear Predictors, Top-Right: Q-Q plot of Quantile Residuals, Bottom: Plot of Cook's Distance for Each Observation.

As has already been stated, a tweedie glm with the log-link function was the final model chosen. In Fig. 4, four diagnostic plots are shown. In the top left is a plot of the quantile residuals against the fitted values transformed by the constant information criteria. In a good model, there should not be any patterns in the plot, however, in this case there is a curved trend. This suggests that the tweedie glm used may not be a good fit for the data. Creating a robust model may require more sophisticated model building techniques.

In Fig. 4 there is also a plot of the working predictors against the linear predictors. This plot is used to check the assumption that the correct link function was used (in this case the log link function was used). The plot shows a linear trend, but no curved trends. This suggests that the link function used is okay.

Figure 4 also shows a Q-Q plot of quantile residuals. The points are quite close to the line which indicates that the normality assumption is satisfied. However, at a theoretical quantile of about 2, the points deviate pretty sharply from the line. For this reason, the normality assumption is believed to be violated.

Table 1 Important statistics from final model: coefficients, standard errors, t-test Statistics, and p-values for each variable

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.2231e+01	5.5017e-02	-222.3170	< 2.2e-16
Disbursed_Amount	7.6842e-04	3.5140e-04	2.1867	0.02879
Original_Principal_Amount	-2.7823e-03	3.2083e-04	-8.6720	< 2.2e-16
RegionEAST ASIA AND PACIFIC	1.7508e-01	2.8512e-02	6.1405	8.573e-10
RegionEast_and_West_Africa	2.4471e-01	3.6309e-02	6.7396	1.686e-11
RegionLATIN AMERICA AND CARIBBEAN	4.2270e-02	2.6385e-02	1.6020	0.10918
RegionOTHER	4.9352e-01	5.0031e-02	9.8642	< 2.2e-16
Loan_StatusActive_(not_fully_repaid_or_cancelled)	-1.6141e+00	6.4897e-02	-24.8713	< 2.2e-16
Loan_StatusFully Repaid	1.2237e-01	5.3526e-02	2.2862	0.02227

Table 1 above shows some important statistics for the final tweedie model that was used to fit the data. All coefficients in this model have p-values less than .05 except for the level Latin America and Caribbean, which has a p-value of .10918. This suggests that countries in Latin America and the Caribbean do not receive interest rates from the IBRD that are significantly different than the reference level: countries receiving loans that are in Europe, the Middle East, North Africa, and Central and South Asia. The decision was made not to combine the variable Latin America and the Caribbean with the reference level because the p-value for the former variable is not extremely high.

It is not very useful to try and interpret the coefficients exactly because the continuous variables were divided by 1,000,000 in the first stages of analysis and because the log link function was used, which distorts the interpretation of how interest rates are changing. However, it is important to note the signs of each statistically significant variable. This shows that the disbursed amount of a loan is positively correlated with the interest rate. It shows that the original principal amount of the loan is negatively correlated with interest rate. This is interesting because one would expect the principal and actual disbursed amount to have similar effects on the response variable. An interesting question that arises is: Why do these two variables have a different effect on the response? Overall, the disbursed amount seems to be more important than the principal amount because the former is how much money a country actually receives.

The coefficients also show that the loan status being active is associated with a lower interest rate than the reference level: which is that a loan has a status of cancelled. This could be because loans with high interest rates are harder to pay off and therefore more likely to be cancelled. In addition, a loan having a status of fully repaid is associated with a higher interest rate than loans with a status of cancelled. This is interesting because it seems to contradict the previous statements.

The coefficients also can be used to infer the effects on interest rates of the variable region. The coefficients for every level of region, except the reference level, are positive.

This means that countries in East Asia and the Pacific, Latin America, East and West Africa, and the category Other, are associated with higher interest rates than the countries in the reference level region: Europe, the Middle East, North Africa, and Central and South Asia.

Section 4 Discussion

The final model chosen for this analysis was a tweedie glm with the link function and an index parameter of 1.2769. While this model performed better than the other models that were tested, the diagnostic plots for the model show that it still had significant problems. Therefore, all conclusions should be considered with extreme caution.

One of the primary goals of this analysis was to determine if the World Bank was engaging in racist lending practices. A firm conclusion about this cannot be made because of significant problems with the model. However, it was interesting to see that all regions that were not included in the reference level were associated with higher interest rates than the reference level. This is interesting because the reference level included regions where, for the most part, a large portion of the population is white. The other regions were areas with larger proportions of people of color. It is also helpful to also note that all countries that are eligible to receive IBRD loans are developing or struggling financially in some way. While a conclusion that the World Bank is engaging in racist lending practices based on this analysis is very premature, it may be an area worth doing more research in.