# Linear Models and Zipf's Law

Khushi Khatri

### Introduction

Linguist George Zipf suggested a functional relationship between word rank and its frequency or rate of usage per 1000 words, and upon taking logs on both sides this relationship evaluates to

$$E[log(f_i)|log(i)] = log(a) - blog(i)$$

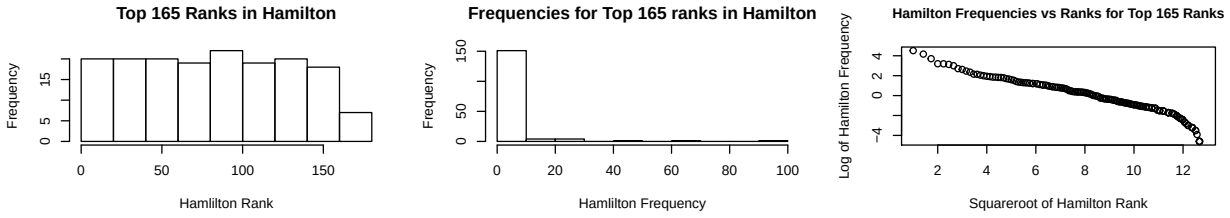
This report uses the data from Hamilton's work in the MWwords dataset in R to investigate Zipf's law.

### Data Description

This dataset, MWwords, is from the R package alr4 (Data to Accompany Applied Linear Regression 4th Edition) that contains data about the number of times certain words appear in the written works of famous authors.

### Analysis

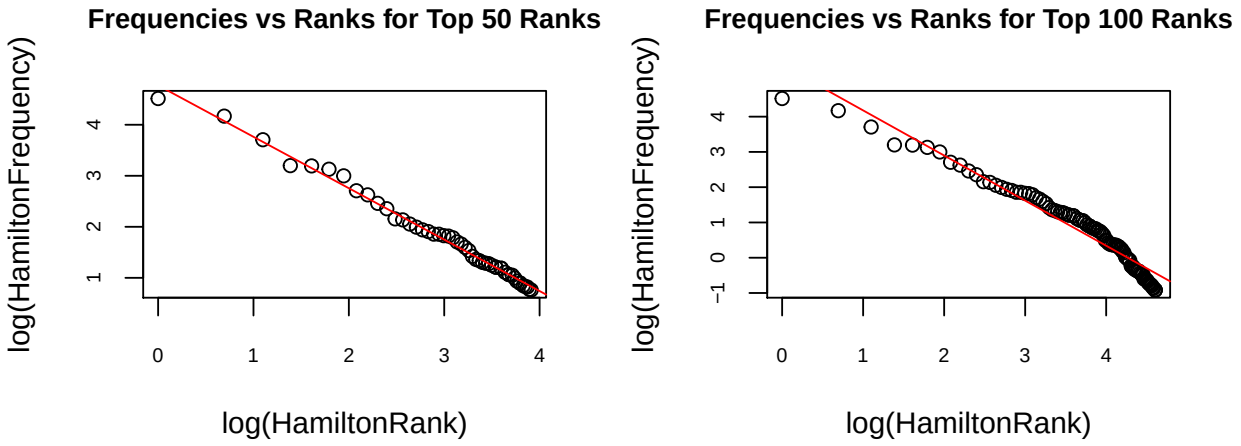
The univariate distributions of Hamilton's ranks and frequencies, and the bivariate distribution of these two variables was examined for the top 165 most frequent words using the graphs below.



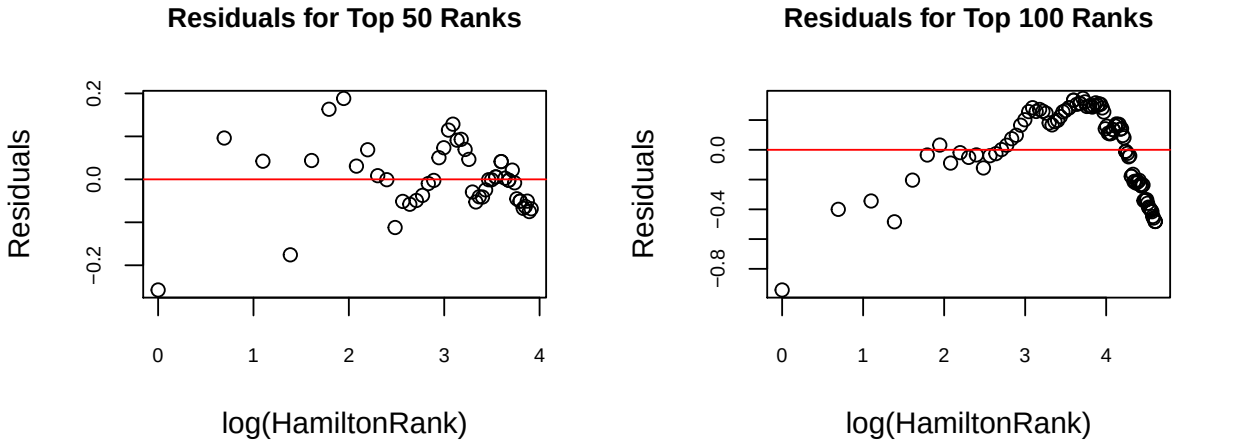

The histogram of Hamilton Ranks reveals that the ranks are roughly uniformly distributed, and hence the distribution is roughly symmetric. Hamilton's frequencies, on the other hand, are extremely right-skewed, with the vast majority of frequency values between 0 and 10 times usage per 1000 words. I initially tried to visualize the bivariate relationship by creating a scatterplot of the two variables. One combination of downward power transformations, taking the log of Hamilton Frequency and the square root of Hamilton Ranks respectively, proved particularly useful in visualizing the data points and linearizing them to increase the appropriate-ness of the linear model. This makes sense because the initial graph of Hamilton Frequencies vs Hamilton Ranks was non-linear but also simple and monotone. According to Mosteller and Tukey's *bulging rule*, I would have to apply power transformations in the downward direction to either or both Hamilton Frequencies and Ranks, which would spread the smaller y and/or x values relative to the larger y and/or x values and linearize the data.

I then created scatterplots of the two variables for the words with top 50 and top 100 ranks respectively. For each of the summary graphs, I have made a scatterplot of the log transformations of Hamilton's frequencies and ranks. I used the log transformation on both axes because the linear model is given after taking

logarithms on both sides of Zipf's law. These transformations also make it easier to visualize the data (as explained by the bulging rule).




The residual plots for the two linear models (top 50 and top 100 ranks) are also shown below.




The co-efficients of the linear model for the top 50 most frequent words data and the predicted average frequency are indicated below:

```
zip=lm(data=top_50, formula=log(Hamilton)~log(HamiltonRank))
zip$coefficients
```

```
##       (Intercept) log(HamiltonRank)
##          4.771236         -1.007639
```

```
predict(zip, data.frame(HamiltonRank=mean(log(top_50$HamiltonRank))))
```

```
##        1
## 3.674504
```

A similar method for the 100 most frequent words yielded an intercept=5.455686 , slope=-1.279090 and a predicted average frequency of 3.467423.

**Discussion**

In the case of the top 50 most frequent words data, the value of b (1.007639) is indeed very close to 1 (as observed by Zipf). The value of b (1.279090) is also close to 1 (although not as close) when the top 100 most frequent words data is considered. Comparisons of the residual plots and scatterplots for the top 50 and top 100 ranks reveals that the regression line is a much better fit for the top 50 ranks data than for the top 100 words data. The residuals in the former plot are somewhat evenly spread about zero. The residuals in the latter plot are not only much larger but also display an evident pattern rather than an even vertical spread throughout the plot. Therefore, with this data, Zipf's law does tend to break down for larger numbers of words and does not predict frequency as well. Since this law tends to break down for this model with larger numbers of words, I think its applicability in causal inference or prediction is very limited and hence it will not be useful for these purposes. Furthermore, the scatterplots and model fits alone can be misleading. I dont think it makes sense to talk about prediction and causal inference in this context (of written works) wherein we describe the frequency of words as a function of their ranks. The regression models provide a summary of the somewhat linear association between log(HamiltonRank) and log(HamiltonFrequency).Therefore I think that the models fit in this exercise are relatively most useful for summarizing association. However, in this application too they have their limitations when larger numbers of words are taken into consideration.

**Conclusion**

An analysis of Zipf's law therefore reveals that it tends to break down for larger numbers of words and is not very useful for prediction or causal inference but can be used to summarise associations.