**Multiple Linear Regression Model Predicting Yelp Rating Using Sentiment Analysis**

Xiangyu Zeng

**Introduction**

Yelp is an Internet company founded in 2004 to “help people find great local businesses”

by providing a platform for users to write reviews of businesses. As of Q4 2016, there are more than 121 million reviews and 89 million unique (average) monthly visitors have visited its website. Our project focuses on sentiment analysis, specifically predicting a reviewer’s rating of a restaurant based on one’s comment.

**Background Information of Data**

The data used to train/predict are extracted from a dataset of 4.1 million reviews. Measurements include name, city, longitude, latitude and category of restaurants, date, stars, usefulness, funniness, coolness, number of characters, number of words and sentiment score (generated using the AFINN lexicon) of comments, and a collection of words with their number of occurrences.

**Motivation for the Model**

In the beginning, we randomly selected 1000 reviews whose ratings were evenly distributed ranging from 1 star to 5 stars from the data provided. We read through each text and listed 68 additional words/phrases apart from the predictors that had already been provided. In addition, we established 16 interaction predictors. Then we divide each word count by number of words of each review. We finally established our model through model selection algorithm (based on good criteria adR2, BIC, Cp, and CV) by combining usefulness, funniness, coolness, and sentiment score of review comments, 25 words that had been provided, 54 newly-added words/phrases, 16 interaction terms, that is, in all, 99 predictors.

**Statement of the Model**

(a) Five observations were removed due to non-English languages.

IDs: 18160 25680 27672 29938 31292

(b) Three observations were removed due to high influential reason.

(i) 16804: high value in Cook's Distance (only one word "best" is included in text)

(ii) 34396: wording (disgusting \* 4, which leads to super negative score)

(iii) 26541: "best", "good", "great" But customer is implicitly expressing a disappointing idea.

**Summary of the Model**

Stars ~ X1 + X2 + X3 + X5 + X7 + X9 + X10 + X11 + X12 + X14 + X15 + X16 + X17 + X18 + X19 + X20 + X21 + X22 + X23 + X24 + X25 + X26 + X27 + X28 + X29 + X30 + X31 + X32 + X33 + X35 + X36 + X37 + X38 + X39 + X40 + X41 + X42 + X43 + X44 + X46 + X47 + X49 + X50 + X52 + X54 + X55 + X56 + X57 + X58 + X59 + X61 + X62 + X63 + X65 + X66 + X67 + X68 + X69 + X71 + X72 + X73 + X75 + X77 + X78 + X80 + X81 + X82 + X84 + X85 + X86 + X87 + X88 + X90 + X91 + X92 + X93 + X94 + X95 + X96 + X98 + X99 + X100 + X101 + X38:X94 + X94:X98 + X71:X75 + X52:X94 + X36:X94 + X36:X52 + X75:X94 + X52:X75 + X36:X75 + X77:X95 + X36:X52:X94 + X52:X75:X94 + X36:X75:X94 + X33:X94:X95 + X33:X77:X95 + X33:X77:X94:X95 (Please see R code for coefficients)

**Interpretation of the Model**

1. Linearity: According to residual plot (Fig. A), there is a clear pattern in general. However, if we look at the middle part, where the stars range from 1 to 5, points are evenly distributed around y = 0.

2. Additivity: Unknown since no interaction plot can be drawn. However, we generated interaction terms using an R package called tree.

3. Constant effects: Reasonable based on wide range of data measurements.

4. Fixed X: Text from customers are directly analyzed.

5. Normality of error: Good linear relationship at 45-degree line of the QQ plot (Fig. B).

**Strengths**

1. Model provides good prediction, with high Kaggle score.

2. Model has an adjusted R-squared value of 0.446

3. Model has an F-statistic (goodness of fit) of 299.3

4. P-Value of F-statistic is less than 2.2 \* 10-16, which suggests that predictors do matter.

**Weaknesses**

1. There are 99 predictors in the model. These 99 predictors could be combined and classified into three categories: positive, negative and sentiment score.

2. Some p-values of some specific predictors are relatively not significant, though they contribute a lot in providing good prediction.

3. All predictions are numeric and discrete (before being divided by number of words). Therefore, points in the residual plot appear in pattern, and we need other packages to help determine interaction terms.

4. If we choose to use other models than Multiple Linear Regression Model, y values could be categorical, which improves prediction even better.

**Conclusion**

Sentiment analysis has become crucial in companies possessing immersive data sets so far. If there were no such analysis, companies would not be able to know what customers think. Therefore, our Multiple Linear Regression Model is a strong tool to implement sentiment analysis. All that is needed for this task is a word/phrase counter and a plain text. By plugging in the 99 predictors in our model, people can simply get a prediction about yelp restaurant ratings.

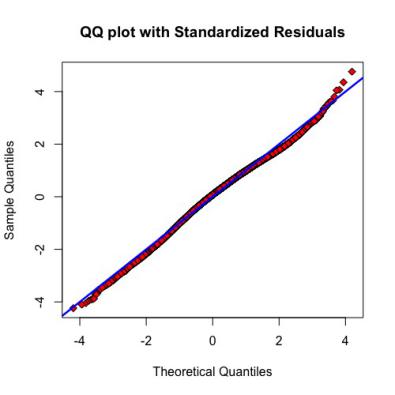
**References**

Package “missForest”, Daniel J. Stekhoven

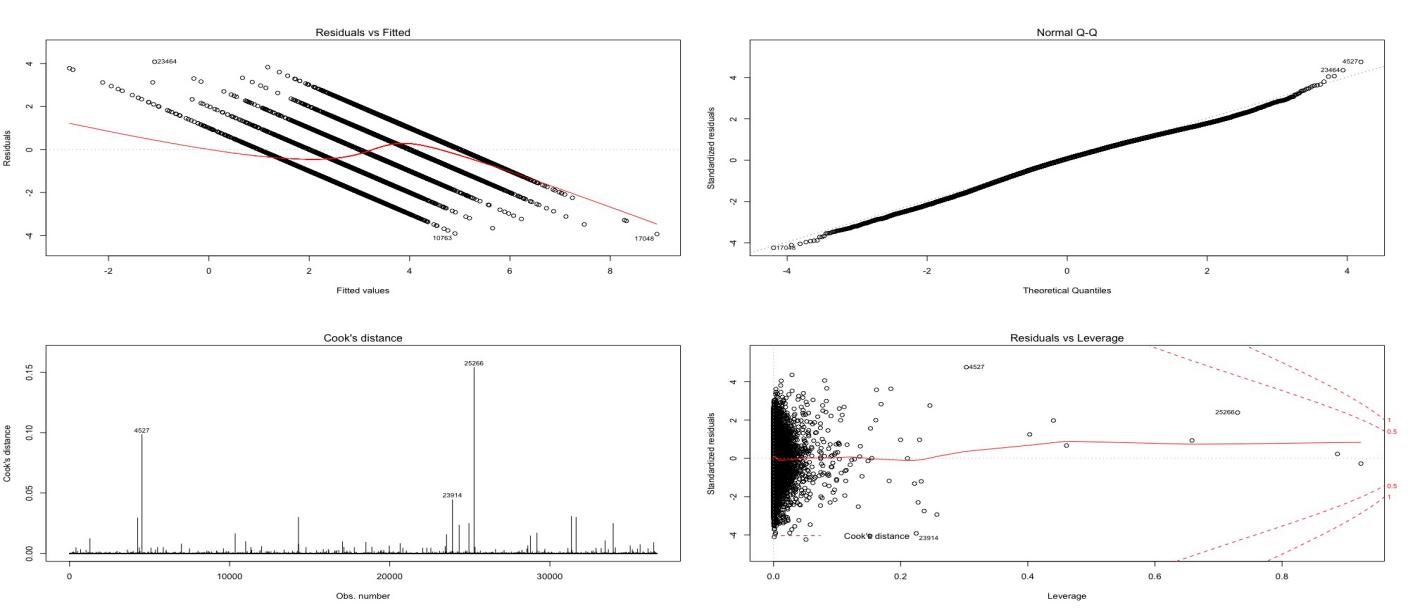
Package “leaps”, Thomas Lumley based on Fortran code by Alan Miller

Package “boot”, Brian Ripley

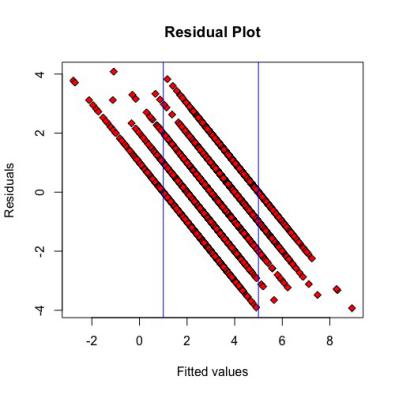
Package “SuperLearner”, Eric Polley



B



A



A