What's in a Rating? Exploring What Factors Influence the Number of Yelp Food Reviews

Winston W. H. Eng

Contents

1 ABSTRACT

1.1 Objective

Yelp is a online service specializing in crowd-sourced reviews for local businesses. In an effort to spur "innovative research", the company has released data set samplings and subsequently encouraged students to dive deep. The objective of this evalution was focused on determining what factors influence the number of reviews a food business receives.

1.2 Methods

1,168 Pittsburgh restaurants were selected from an international cohort. A number of user-submitted factors such as perceived price and quality of meal were collected alongside the restaurants' services such as availability for takeout and reservations. Simple linear regression was utilized to assess the relationship between these independent variables and the number of reviews a restaurant received. A final prediction model was created using information determined from a spearman ρ rank squared test, and its validity was tested against automatic processes such as Stepwise Regression and LASSO.

1.3 Results

Aside from one variable, each of the predictors, individually, had a statistically significant ability to predict the number of reviews a restaurant would receive; related r^2 values ranged from 0.02 to 0.15 with one variable having as high a value as 0.516. In contrast, the final prediction model (dubbed the "Spearman Regression Model") had the highest r^2 (0.801) and lowest MSE (0.267), AIC (1832), and BIC (1984) values. The "Stepwise Regression" and "LASSO" techniques performed no better than the "Spearman Approach"" in all four of those categories.

1.4 Conclusion

Not all of the predictors exhibited monotonic relationships with the outcome variable. The "Spearman Approach" was created as a consequence of putting this theory into practice and manually choosing where to spend the degrees of freedom. The resulting model produced the most accurate results and subsequently provided greater insights into which predictors best influence the number of reviews a food business would receive.

2 INTRODUCTION

Electronic word-of-mouth (eWOM) can be defined by "any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet." (Hennig-Thurau et al. 2004) It has often been credited as having an incredibly amount of influence in our current culture; customers have been cited as seeing online reviews as "more trustworthy and persuasive than traditional media, such as print ads, personal selling, and radio and TV advertising" and "more influential in their decision than speaking with friends in person". (Cheung and Thadani 2012, Steffes and Burgee (2009)) It should be no surprise that engaging in such online communication spectrums could be seen as extremely beneficial in helping to assess the wellbeing of a current business or product.

Since 2004, the world has had Yelp, a website and mobile app designed around the concept of rating business, culinary and otherwise, in a numerical and descriptive manner. With 24 and 73 million unique users on mobile and desktop respectively, the company fits into the category of "Web 2.0 sites" where user-submitted reviews are targeted at a primary audience of other consumers. (Yelp 2017, Tucker (2011)) In the words of Stephanie Ichinos, ex-Senior Director of Communications, these reviews "supply information to the bloggers, support local businesses, [and] help share potentially useful information with others…" These criticisms are heavily integrated into the livelihood of these establishments; even an "one-star increase in Yelp rating" has been shown to "lead to over a 5-9 percent increase in revenue". (Luca 2016)

In an effort to contribute to the literature surrounding the influence of eWOM on local food establishments, this study seeks to determine what combination of restaurant factors impact the number of reviews it will receive. From an initial assessment of over 4.1 million reviews and 947,000 tips by 1 million users for 144,000 business provided by the Yelp in the 9th round of the Yelp Dataset Challenge, an objective was developed to create a prediction model capable of relating the relationships amongst these variables most accurately. Given the complexity and difficulty assumed in relating qualitative experiences with quantitative measures, the author hypothesizes that there might exist non-monotonic relationships within the data which may influence the considerations taken during model generation.

3 METHODOLOGY

3.1 Sample Selection

The Yelp Dataset included 144,072 businesses spanning 11 international locations. Businesses not within the food industry were excluded. Additionally, restaurants that were missing data on any of the variables were excluded. In total, the final sample size consisted of 1,168 currently open restaurants within the city of Pittsburgh.

3.2 Data Dictionary

Table 1: Data Dictionary of Regression Model Variables

Variable.Name	Description	Variable.Type
stars	Number of Stars a Restaurant Received	Predictor
bar	Type of Alcohol Service Available	Predictor
Takeout	Availability of Takeout	Predictor
Delivery	Availability of Delivery	Predictor
Reservations	Availability to set a Reservation	Predictor
GF.breakfast	Considered 'Good For Breakfast'	Predictor
GF.brunch	Considered 'Good For Brunch'	Predictor
GF.lunch	Considered 'Good For Lunch'	Predictor
GF.dinner	Considered 'Good For Dinner'	Predictor
GF.dessert	Considered 'Good For Dessert'	Predictor
GF.latenight	Considered 'Good For Late Night'	Predictor
tips	Total Number of Tips a Restaurant Received	Predictor
review_count	The Number of Reviews a Restaurant Received	Outcome

For ease and consistency, the variables will be **bolded** and referenced using the names within the data dictionary. Parenthesized superscripts will act as references to graphics within the Appendix section of the paper.

3.3 Transformation of Outcome Variable

The outcome variable **review_count** did not follow a normal distribution $^{(6.1)}$. After applying a log transformation to the data, it appeared more normally distributed $^{(6.2)}$, so this form of **review_count** was used in the analyses. These assessments took the form of kernel density plots.

3.4 Statistical Analysis

Descriptive statistics included assessment of the number, mean, and standard deviation of the different categorical variables. Additionally, Violin plots provided a visual assessment of the distribution of **review_count** per each variable; kernel density estimates were mirrored and formed a symmetrical shape. Red points indicated outliers or values outside 1.5 times the interquartile range, while white points demonstrated the median value; the black boxes acted as boxplots. To test linear relationships, scatterplots were used, and collinearity was evaluated via variance inflation factor calculations.

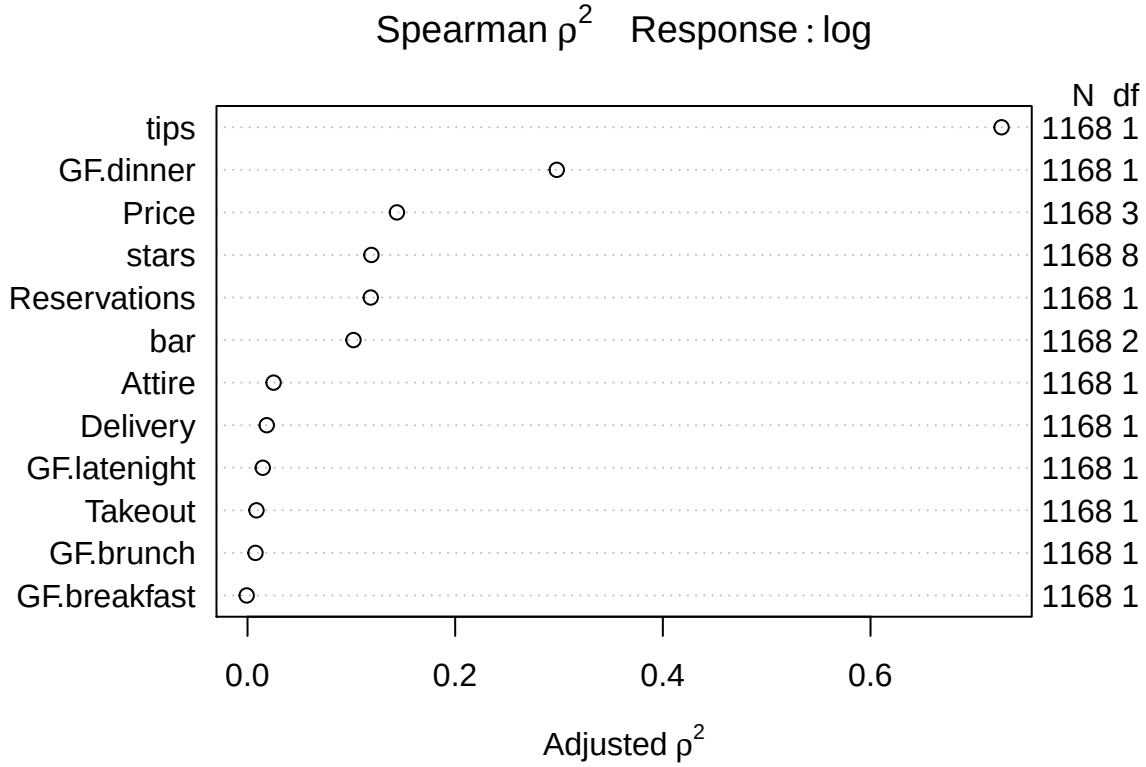
Initially, a simple linear regression was used to determine how well each independent variable would be able to predict **review_count**; if the f-statistic had a related p-value less than a critical p-value of 0.05, the predictor was considered to have a statistically significant relationship with the outcome. While interactions were considered amongst all of the different variables, it was worth noting that to aid with interpreting and maintaining a more parsimonious model, each possible iteration was not considered.

Subsequently, two different model approaches were used: 1) Manual vs 2) Automatic. As the objective was created with the former process in mind, the automatic techniques offered an opportunity to test the manual model's validity.

An assumption used was that each of the independent variables had a monotonic relationship with the outcome variable. However, if this were to be false, the ability for a model to accurately assess the outcome variable could be called into question. Therefore, in order to confirm the potential existence of non-monotonic relationships, evaluation of the square of a quadratic rank generalized of Spearman's ρ coefficient was necessary.

For the first category, a Spearman's ρ^2 was initially ran to determine if all relationships between each of the independent variables with the outcome were monotonic. In general, the spearman ρ^2 will "detect not only nonlinear relationships (as will ordinary Spearman ρ) but some non-monotonic ones as well." (Harrell 2015)

It has been suggested that there shouldn't be more than 15 observations per predictors (Harrell 2015); while, others suggests a maximum of 10 observations per predictor. (Vittinghoff et al. 2011) Since there were 1168 observations, the max amount of predictors that could theoretically be kept within the model would be 77. As there were already 12 variables within the equation, it was deemed acceptable to allow for transformations of the variables without violating these regression guidelines.



Based on the Spearman Plot, it appeared that the three most likely to have a non-monotonic relationship with the outcome variable were **tips**, **GF.dinner**, and **Price**.

The **tips** variable had a quartic spline applied to it as well as interactions set between itself and **GF.dinner** and **Price** respectively. Justification is that, by convention, a greater number of tips should

indicate an overall greater "perception" of the restaurant; this should hold especially for dinner where special occasions are generally more common. Whether positively or negatively, if a restaurant had enough of an impact on an individual as to have that person leave a tip, it is very possible to expect that same customer to leave a review as well. Furthermore, restaurants that typically stand out to people are those that could be described as "worthwhile experiences". For a college student, this may mean a filling, delicious, and cheap meal, while for a person looking for a fancy dinner, it may mean splurging on a place with a reputation to be upscale. Demographics of Yelp aside, it should be possible to imagine that **tips** and **price** could have a meaningful interaction in such a manner. Subsequently, after this spearman ρ assessment, an ordinary least squares (OLS) regression model was then created from these selected variables.

For the second category, bi-directional stepwise regression and Least Absolute Shrinkage and Selection Operator (LASSO) models were run. The stepwise approach utilized an AIC criteria that dropped or added variables based on $-2 \times (loglikelihood) + 2p$ where `p` was the rank of the model, while the LASSO relied on the least angle regression algorithm developed by Tibshirani, *et al.* (Tibshirani 1996)

Following the creation of the models, all of the comparative outputs were summarized in Table 2. Furthermore, studentized residuals were calculated and plotted against the fitted values, while histograms were created to determine their overall normality. Leverage and Cook's Distance plots were created to determine outliers and influential points; if points were deemed statistically significantly influential, the models were assessed with and without those points' inclusion. All analysis and discussion was created using RStudio's `RMarkdown` and `R Version 1.0.136`.

4 Results

4.1 Descriptive Statistics

When looking at the descriptive statistics of each of the different variables, it is worth noting that the distribution of **review_count** is not equal amongst all of the different levels within the 13 independent variables. For instance as demonstrated in Reference 6.3, the distribution of **stars** is not equal; 12 restaurants had a 1.5 star rating, while 353 had 4 stars.(6.3) This is also physically evident in the subsequent graphic comprising of violin plots.(6.4) It is worth noting that most of the variables used were categorical and had between 2 to 3 levels. When looking at the scatterplots of each independent variable against the **review_count**, most appeared linear; however, the scatterplot for **review_count** vs **stars** appeared to be curved.(6.5)

4.2 Comparing Regression Models

Table 2: Regression Models and their Outputs

Model	R.Squared	MSE	F.Statistic	P-Value	AIC	BIC
stars	0.127	1.172	21.07	< 2.2e-16	3520	3571
bar	0.101	1.207	65.23	< 2.2e-16	3543	3563
Takeout	0.009	1.331	10.18	0.001	3654	3670
Delivery	0.018	1.318	21.35	< 2.2e-16	3643	3658
Price	0.141	1.153	63.81	< 2.2e-16	3491	3516
Attire	0.024	1.31	28.41	< 2.2e-16	3636	3652
Reservations	0.106	1.2	138.2	< 2.2e-16	3534	3549
GF.dessert	0.004	1.337	4.443	0.035	3660	3675
GF.latenight	0.014	1.324	16.35	< 2.2e-16	3648	3663
GF.lunch	0.037	1.293	44.34	< 2.2e-16	3621	3636
GF.dinner	0.281	0.965	455.4	< 2.2e-16	3279	3295
GF.breakfast	0	1.342	0.066	0.798	3664	3680
GF.brunch	0.01	1.329	11.28	0.001	3653	3668
tips	0.516	0.65	1242	< 2.2e-16	2817	2832
Spearman Approach	0.801	0.267	163.8	< 2.2e-16	1832	1984
Stepwise Approach	0.679	0.43	121.6	< 2.2e-16	2374	2485
LASSO Approach	0.68	0.429	110.7	< 2.2e-16	2375	2496

4.3 Simple Linear Regression Models

For the simple linear regression models, each independent variable, aside from **GF.breakfast**, was statistically significantly able to predict the **review_count** against a critical p-value of 0.05. For the most part, the amount of variation for which each variable was able to account was between 2-15%. An anomaly, the **tips** variable demonstrated the highest r^2 value of 0.516. The related q-q plots showed no deviations from normality; this was also true for the **GF.breakfast** model. However, the studentized vs fitted residuals graphs did generally have fanning.(6.6)

4.4 Spearman Regression Model

After creating the Spearman ρ Rank Square Plot and selecting to spend degrees of freedom on the **price**, **GF.dinner**, and **tips** variables, the model included the following variables:

Table 3: Formula of Spearman Regression Model with Notes (continued below)

Model	Formula
Spearman Approach	log ~ rcs(tips, 4) + stars + bar + Takeout + Delivery + Attire + Reservations + GF.brunch + GF.breakfast + GF.latenight + GF.dinner + Price + GF.dinner %ia% tips + Price %ia% tips

Notes
'rcs(x, 4)' = quartic spline on the variable 'x', 'y %ia% z' = interaction between variable y and z

A quartic spline was applied to the **tips** variable, while interactions took place between **tips** and **GF.dinner** and **Price** respectively.

With a F-statistic of 163.8 and a p-value < 0.001, the spearman model was able to account for over 80.1% of the variation; it had the lowest MSE, AIC, and BIC values of 0.267, 1832, and 1984 respectively. When assessing the variance inflation factor (VIF) of the variables, there were some that had values greater than the a critical value of 5.0. However, these outliers tended to fall within two categories, either 1) they were indicator variables representing categories with 3 or more levels or 2) they were interactions between different variables. Having high VIFs within categorical dummy variables should not be considered problematic as often "nothing else in the regression is affected." (Allison 2012) Additionally, higher VIF values are expected when including power transformations or interactions of other variables; by their very natures, there should exist influential underlying relationships.

When plotting the residuals vs fitted, there was a very slight funneling in the residuals vs fitted plot, possibly demonstrating some non-linear pattern within the residuals. [(6.7)] However, in the scale-location plot, the residuals appeared equally spread out, indicating no homoscedasticity violation. Additionally, the Q-Q plot appeared to be normal, and there did not appear to be any points with high influence or leverage.

4.5 Stepwise Regression Model

Recall that the stepwise regression approach is helpful when trying to determine, from a handful of variables, which best combination can lead to the lowest AIC or lowest amount of information loss.

Table 5: Formula of Stepwise Regression Model with Notes

	Model	Formula	Notes
2	Stepwise Approach	log ~ stars + bar + Price + Attire + Reservations + GF.brunch + GF.breakfast + GF.latenight + GF.dinner + tips	NA

With a F-statistic of 121.6 and a p-value < 0.001, the spearman model was able to account for over 67.9% of the variation; its MSE, AIC, and BIC values were 0.43, 2374, and 2485 respectively. Unlike the spearman regression, this model did not retain the **Takeout** or **Delivery** variables. However, it's VIF levels mirrored similar hightened results for the same variables as the spearman regression had produced. When

plotting the residual versus fitted, there was an intense clustering and funneling of the data.

4.6 LASSO Regression Model

In theory, the LASSO approach is most useful when it is believed that not all of the predictors are having an incredibly impactful effect on predicting the outcome. When running the model, none of the coefficients, after cross-validation, were dropped. Hence, via the LASSO approach, it is suggested that all of the different variables should be retained within the prediction model. Compared to the Spearman Regression, LASSO lacked the interacions and quartic spline.

Table 6: Formula of LASSO Regression Model with Notes (continued below)

	Model	Formula
3	LASSO Approach	log ~ stars + bar + Takeout + Delivery + Price + Attire + Reservations + GF.brunch + GF.breakfast + GF.latenight + GF.dinner + tips

	Notes
3	the same as the full model

With a F-statistic of 110.7 and a p-value < 0.001, the LASSO model was able to account for over 68% of the variation; its MSE, AIC, and BIC values were 0.429, 2375, and 2496 respectively. It's VIF and residuals vs fitted plot mirrored that of the stepwise regression model.

5 Discussion

The main focus of this study was to determine which best combination of factors was most capable of accurately determining the number of reviews a restaurant would receive. In this case, I found that the spearman regression approach produced the most accurate model when compared to the stepwise and LASSO techniques. While, the model could only account for 80.1% of the variation, it outperformed all other approaches, demonstrating that non-monotonic relationships should be included and assessed when dealing with this type of data.

With regards to the regression coefficients, each of them were statistically significant against a critical p-value of 0.05 except for 1) the indicator variable regardng whether a restaurant was "good for breakfast" ($p = 0.121$) and 2) the interaction between the perceived price of a restaurant and the number of tips it received ($p = 0.464367$).

With regards to the **GF.Breakfast** variable, this type of non-statistical significant contradicts what may be found within the stepwise and LASSO models; the same variable had critical p-values < 0.001 in their results. This type of statistically significant reversal happened for the **Takeout** variable as well. While the stepwise model dropped **Takeout** and the LASSO model deemed it non-significant, the Spearman Regression model's version had treated it as statistically signficant with a p-value of 0.00335.

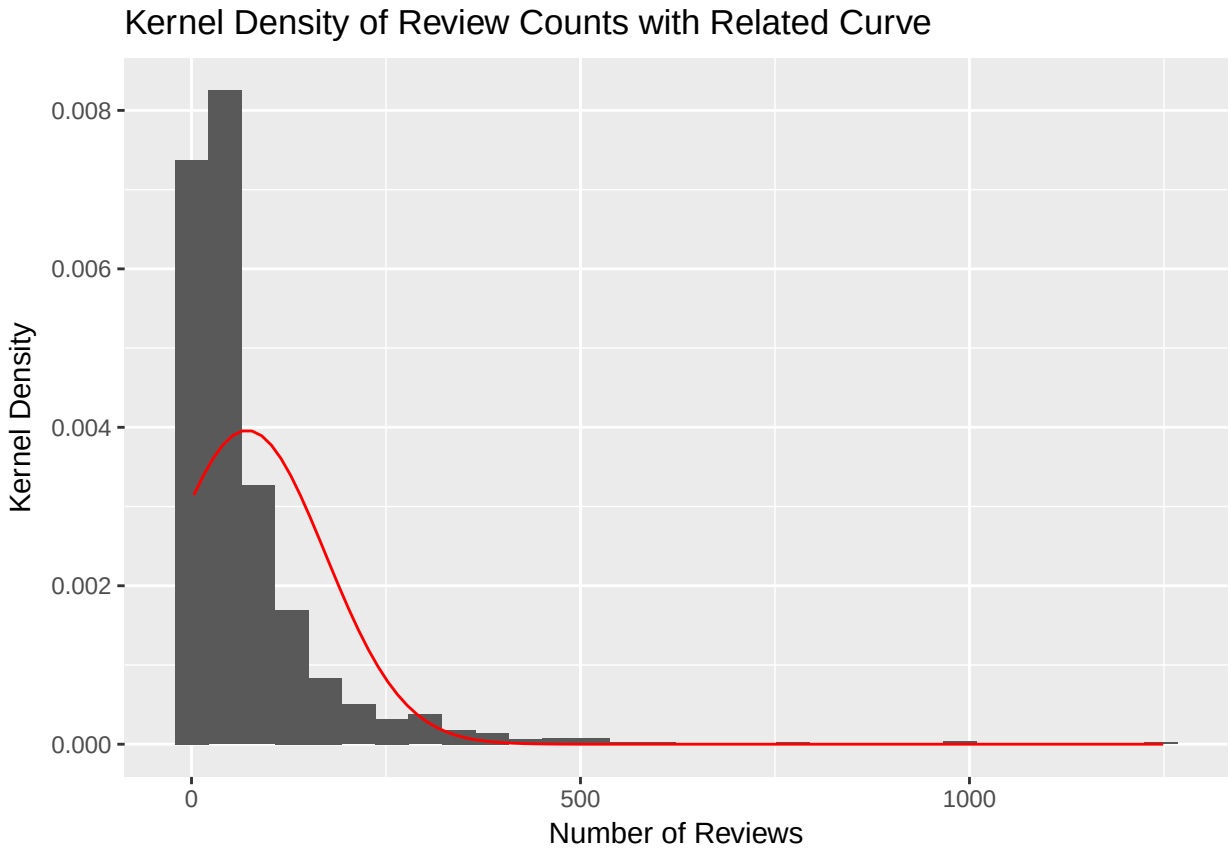
Surprisingly, this appears to coincide with previous research which rated "takeout" as the second most cared about aspect of a restaurant based on Yelp reviews. (Huang, Rogers, and Joo 2014) Additionally, Huang *et al.* found that "breakfast" was rated on the lower end of importance, being a part of only 0.59% of all reviews.

Limitations to this analysis sources heavily from the failure to include additional data types provided within the Yelp Dataset Challenge. While the author did account for the number of reviews a specific restaurant received, there was not a deeper dive into the specific words used within the reviews themselves; perhaps, appling a variation of natural language processing could have given greater insights into the quality of reviews and how that would influence the type of reviews a restaurant would receive. Moreover, visual pieces such as pictures of food, restaurants themselves, and even the patrons/reviewers were not evaluated. Peak times and the cyclic nature of businesses during the workshift were not included; as previous analyses have cited "service"" and "decor" as most important "hidden topics" for reviewers when looking at a restaurant, future work should consider including these types of variables into the model. (Huang, Rogers, and Joo 2014)

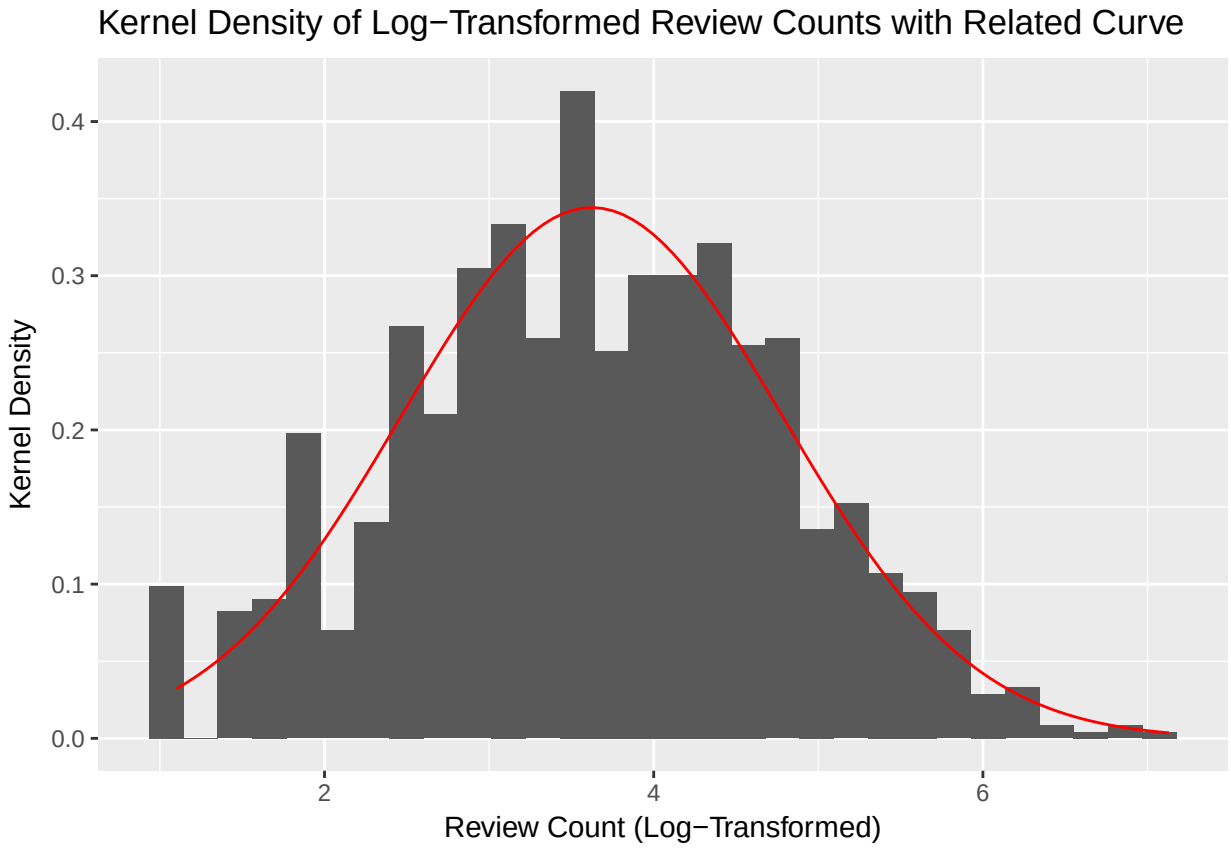
It is clear that there are environmental factors that were not considered in this specific analysis. However, given the assumptions that were made within this paper, there doesn't appear to be an irregular deviation from what has been generally discussed in the literature.

6 Appendix

6.1 Kernel Density Distribution of Review Count



6.2 Kernel Density Distribution of Log-Transformed Review Count

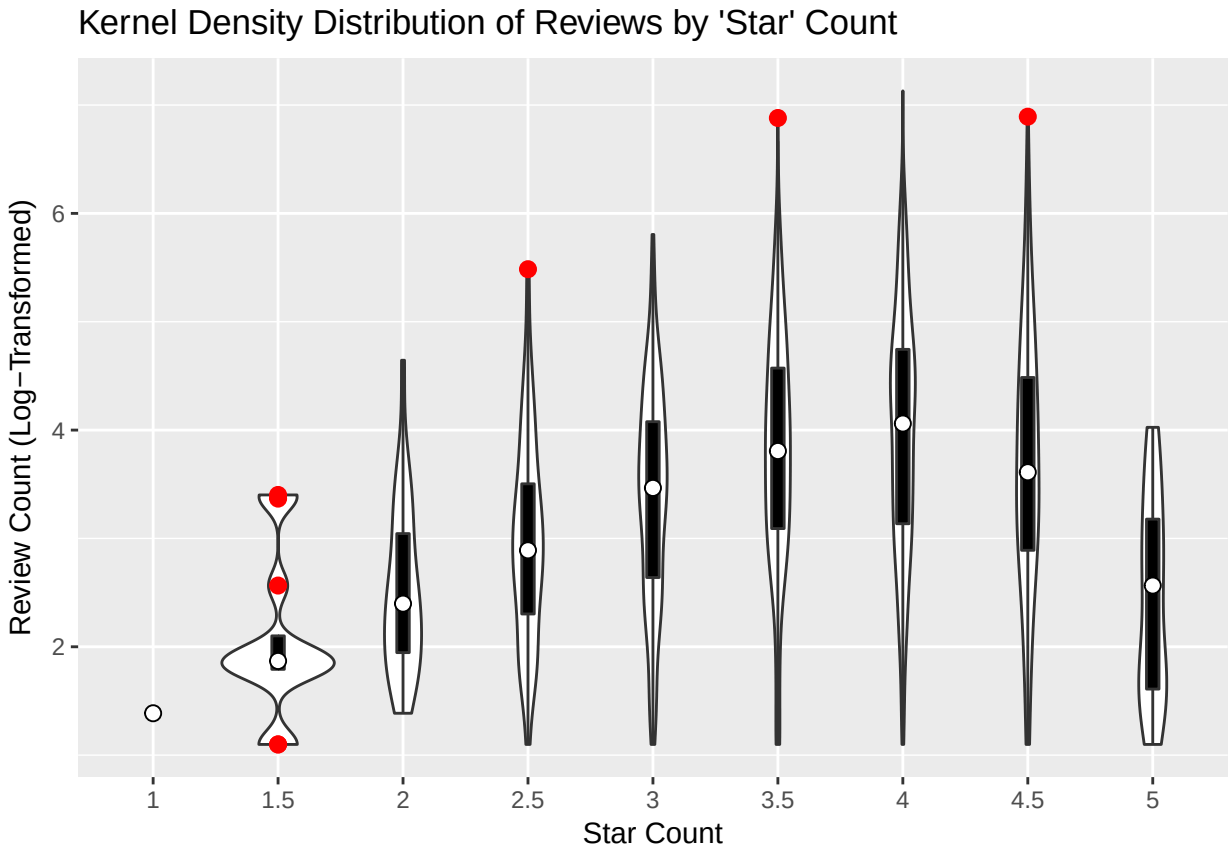


6.3 Summary Statistics of Stars Count (N, Mean, Standard Deviation)

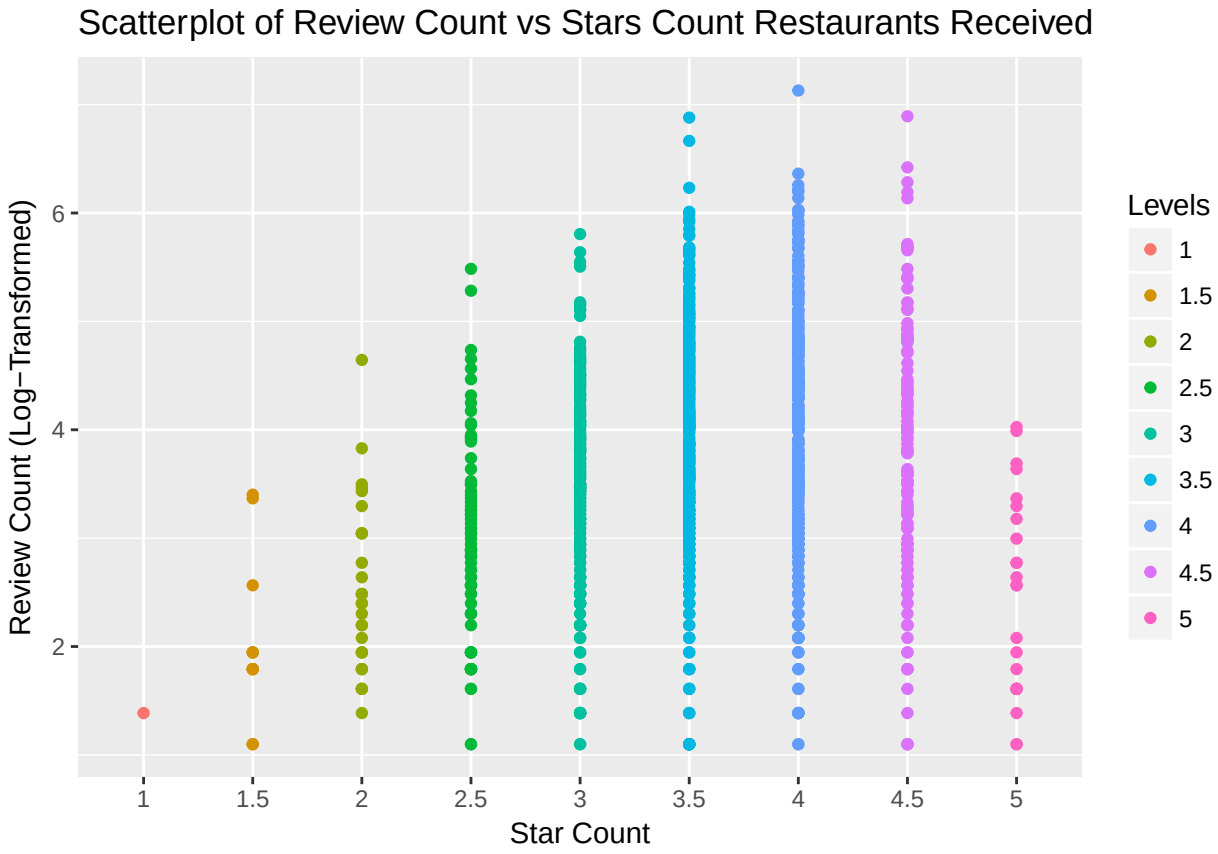
Table 8: Distribution of Stars Assigned to Restaurants

stars	N	Mean	Standard.Deviation
1	1	1.386	NA
1.5	12	2.045	0.735
2	30	2.505	0.771
2.5	76	2.983	0.944
3	201	3.386	0.996
3.5	338	3.783	1.11
4	353	3.953	1.13
4.5	132	3.712	1.223
5	25	2.422	0.919

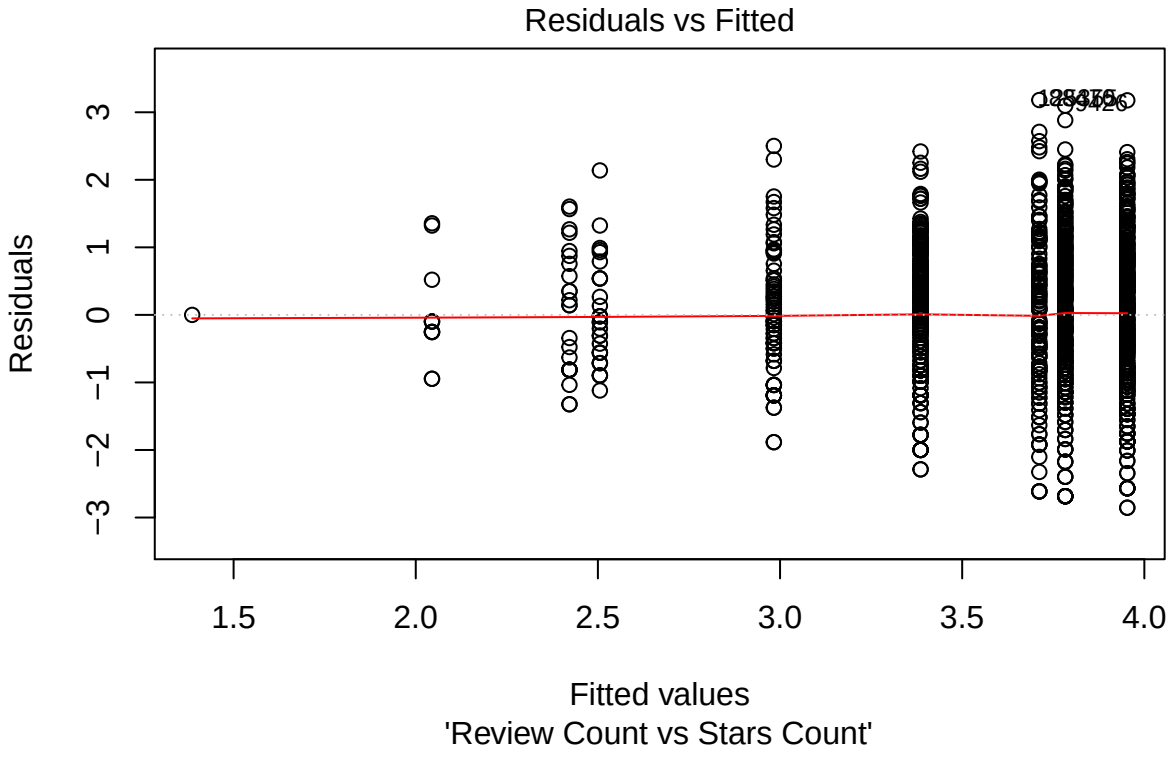
6.4 Kernel Density Distribution of Stars Count via Violin Plots



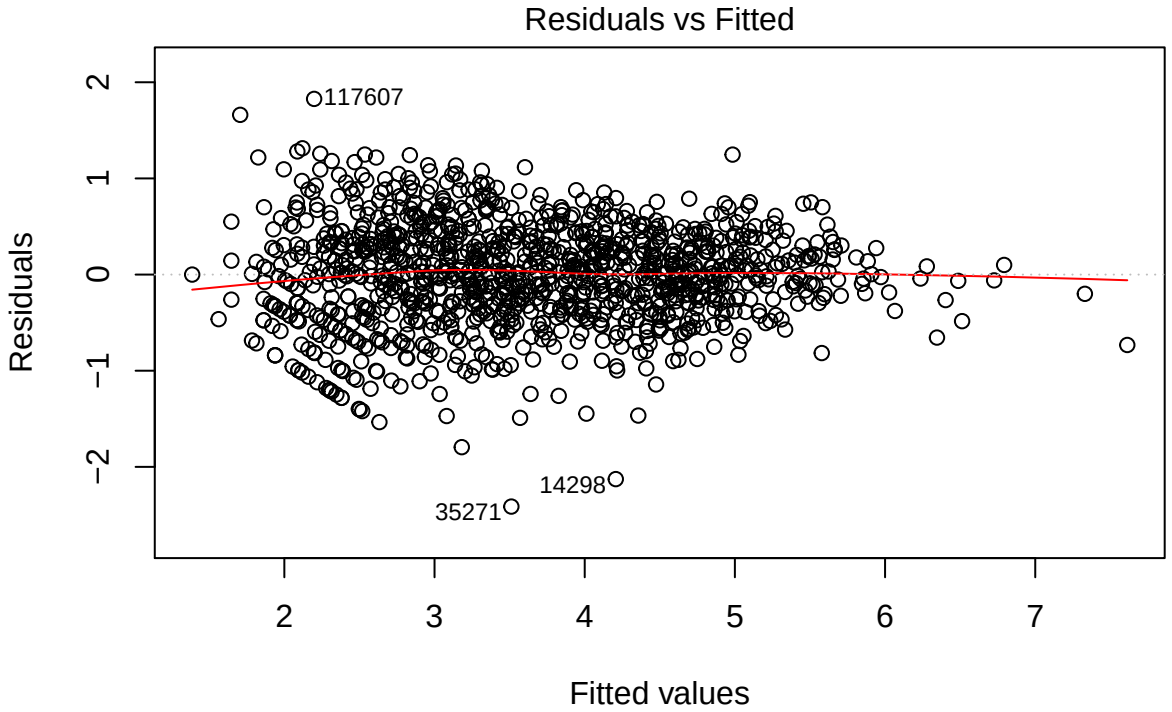
6.5 Scatterplot of Review Count vs Stars Count



6.6 Studentized Residuals vs Fitted Plot for Stars Count Linear Regression



6.7 Studentized Residuals Vs Fitted Plot for Spearman Regression Model



References

Allison, Paul. 2012. "When Can You Safely Ignore Multicollinearity?" Statistical Horizons.

Cheung, Christy MK, and Dimple R Thadani. 2012. "The Impact of Electronic Word-of-Mouth Communication: A Literature Analysis and Integrative Model." *Decision Support Systems* 54 (1). Elsevier: 461–70.

Harrell, Frank. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer.

Hennig-Thurau, Thorsten, Kevin P Gwinner, Gianfranco Walsh, and Dwayne D Gremler. 2004. "Electronic Word-of-Mouth via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet?" *Journal of Interactive Marketing* 18 (1). Elsevier: 38–52.

Huang, James, Stephanie Rogers, and Eunkwang Joo. 2014. "Improving Restaurants by Extracting Subtopics from Yelp Reviews." *IConference 2014 (Social Media Expo)*. iSchools.

Luca, Michael. 2016. "Reviews, Reputation, and Revenue: The Case of Yelp. Com."

Steffes, Erin M, and Lawrence E Burgee. 2009. "Social Ties and Online Word of Mouth." *Internet Research* 19 (1). Emerald Group Publishing Limited: 42–59.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 267–88.

Tucker, Tiana. 2011. "Online Word of Mouth: Characteristics of Yelp. Com Reviews." *Elon Journal of Undergraduate Research in Communications* 2: 37–42.

Vittinghoff, Eric, David V Glidden, Stephen C Shiboski, and Charles E McCulloch. 2011. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer Science & Business Media.

Yelp. 2017. "About Us 10 Things You Should Know About Yelp." https://www.yelp.com/about.