Facebook Post Comment Volume Regression Analysis

Ta-Hung (Denny) Chen

Department of Mathematics and Statistics

Boston University, Boston, Massachusetts, U.S

Instructor: Dr. Fotios Kokkotos

December, 12, 2023

Abstract

In the dynamic realm of social media, the volume of comments a Facebook post garners serves as a crucial indicator of its engagement and reach. This study delves into the predictive factors influencing comment volume, utilizing the Facebook Comment Volume Dataset from the UCI Machine Learning Repository. Our research leverages mixed effect model to construct a negative binomial regression that predicts the number of comments a post is likely to receive within the subsequent hours of its publication. Drawing on the work of Kamaljot Singh and others, I employed count data regression methods and its extensions to account for potential overdispersion. By examining various post features, such as page characteristics, essential and weekday features, and other basic attributes, I endeavor to identify the key determinants of comment volume. The model's accuracy will be assessed using Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), alongside Posterior Predictive Check plots. Our findings aim to empower content creators and social media strategists to amplify their online presence and foster organic user interactions effectively. By providing insights into the promotion of Facebook posts without relying on paid advertising, this research seeks to democratize the approach to enhancing social media visibility.

Introduction:

The study focuses on forecasting Facebook comment volumes using mixed effect regression models, essential for gauging social media engagement. As online interactions dominate today's social sphere, predicting user engagement on Facebook is key for creators and marketers. This research utilizes the Facebook Comment Volume Dataset by Singh and Kaur (2015) to analyze the comments a post receives in the first three days—critical for assessing user interaction. The dataset provides features related to the posts for a detailed examination of the factors affecting comment volumes. Count data regression modeling, starting with Poisson regression, is the methodology used, with flexibility to adapt to other models in case of data overdispersion. The research aims to identify factors that significantly affect Facebook comment volumes and to test the predictive performance of the models. The findings are intended to enhance social media engagement strategies. In essence, this paper advances social media analytics by applying a hierarchical model to predict and understand Facebook user engagement.

Datasets:

The analysis conducted in this research is based on data sourced from the Comment Volume Prediction using Neural Networks and Decision Trees, originally collected by Singh and Kaur (2015) and made available through the UCI Machine Learning Repository. The dataset originates from Facebook Pages and has been meticulously prepared to facilitate the study of comment volume on posts. According to the authors of the research that produced the data, it is presumed that only comments posted within the last three days relative to a given Base date/time[1] are relevant, as older posts are not typically expected to gain further engagement.

To ensure data integrity, any posts lacking comments or other essential information have been excluded. The dataset is divided into two parts: training and testing. The training data encompasses post information collected at five distinct time intervals, resulting in five different data variants.[2] The fifth dataset, known as *Data Variant 5*, has been selected for analysis due to its comprehensive nature and the richness of its observations. Regarding the testing dataset, it comprises 10 test cases, each with 100 observations, and is merged together.

Predictors contain (1) Page features, (2) Essential features, (3) Weekday features, (4) Other basic features. Followed the feature definition by Singh and Kaur (2015).

(1) Page features:

Four features of the category were identified to define the characteristic of the post. *Page likes:* It is a feature that defines users support for specific comments, pictures, wall posts, statuses,

[1] Base date/time is selected to simulated the scenario, as we already know what will happen after this. There is one more kind of time we used in this formulation: is the post published time, which comes before the selected base date/time. See appendix 2 for more information. Singh and Kaur (2015)

[2] Data variants are different samples that are collected at different Base date/time. Singh and Kaur (2015) Figure 3.

or pages. *Page Category:* This defined the category of source of document eg: local business or place, brand, or product, company or institution, artist, band, entertainment, community, etc. *Page Check in's:* The feature shows the presence of the post at particular place. *Page Talking About:* The actual count of users who are "engaged" and interacting with the Page. Including the activities such as comments, likes, shares.

(2) Essential features:

Essential features indicate the pattern of comments on the post within various time interval with reference to random select Base date/time. *CC1:* Total comment count within 72 hours before the selected Base date/time. *CC2:* Comment count in last 24 hours w.r.t to the selected Base date/time. *CC3:* Comment count between last 24 hours to last 48 hours w.r.t to the selected Base date/time. *CC4:* Comment count in first 24 hours w.r.t the selected Base date/time. *CC5:* the difference between *CC2* and *CC3*. And the data also contains min, max, standard deviation, median, and mean of *CC1* to *CC5*.

(3) Weekday features:

Weekday features represent as a binary indicator showing the day on which the post was published and the day on selected Base date/time.

(4) Other basic features:

Other basic features show additional information of the post, including the length of the document, time gap between selected Base date/time and document published ranges from [0,72], document promotion status, and post share count.

Methodology:

The problem planned to be addressed using a count data regression predictive modeling approach, with Poisson regression initially employed. There is flexibility to modify the regression assumptions to better align with the data and yield accurate results. For instance, should overdispersion be detected within the data, alternative models such as Quasi-Poisson or negative binomial regression may be utilized.

- Potential assumption adjustment for Poisson model:

  - Quasi-Poisson regression:

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \cdots$$

Where:

μ: Expected value of the observation i.

$\beta_0, \beta_1, \beta_2, \ldots$: Coefficient to be estimated.

$X_{1_i}, X_{2_i}, \ldots$: Predictors' value for observation i.

Variance for Quasi-Poisson:

$$var(Y) = \mu \cdot \phi, \qquad \phi > 1$$

ϕ is the dispersion coefficient. While $\phi = 1$ is a Poisson distribution with assumptions of parameters $\lambda = \mu = var(Y)$.

- Negative binomial regression:

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \cdots$$

Where:

μ: Expected value of the observation i.

$\beta_0, \beta_1, \beta_2, \ldots$: Coefficient to be estimated.

$X_{1_i}, X_{2_i}, \ldots$: Predictors' value for observation i.

Variance for negative binomial regression:

$$var(Y) = \mu + \alpha\mu^2 = \mu + \frac{\mu^2}{k}$$

α, k are the dispersion parameter in different presenting form. While $\alpha = 0$, or $k \to \infty$, negative binomial model converges to a Poisson distribution with assumptions of parameters $\lambda = \mu = var(Y)$.

- Zero inflation model:

Since overdispersion is often observed in counting data problem, a zero-inflation negative binomial regression will be adopted if required. It combines two parts, the count model and the zero-inflation model.

1. Count Model (Negative Binomial Part): Same as the description of negative binomial regression above.

2. Zero-inflation Model:

$$logit(\pi_i) = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \cdots + \alpha_l Z_{il}$$

Where:

π_i: is the probability of i-th observation is an "extra" zero.

$Z_{i1}, Z_{i2}, \ldots, Z_{il}$: the independent variables for the zero-inflation part.

$\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_l$: the coefficients for the zero-inflation part.

Validation:

Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) will be used to evaluate the accuracy for the regression problem.

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y - \hat{y})^2$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Posterior Predicting Check plot are planned to be considered to evaluate the fitness of the regression model.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(y - \hat{y})^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{N}|y - \hat{y}|$$

Quantitative Analysis:

Exploratory Data Analysis:

Most of the data cleansing and manipulation work had been done by Singh and Kaur (2015), and I directly started the analysis based on the cleaned and well-organized data which is available through UCI-Machine Learning Repository. Refer to what I have mentioned in the Datasets section is that *CC4* is the comment volume in the first 24 hours, and *CC2* is the

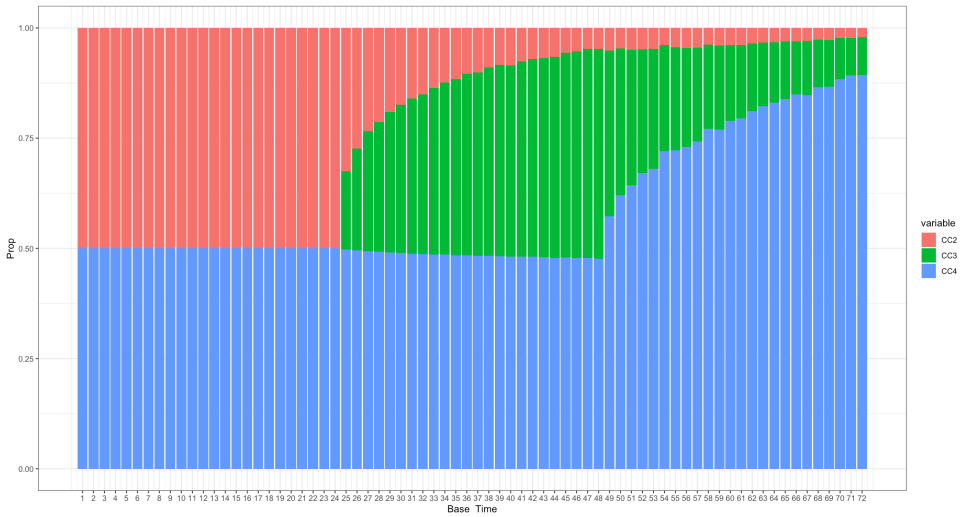


Figure 1. Proportion of CC2, CC3, CC4 to Base_Time

comment volume in last 24 hours w.r.t to the selected Base date/time, which means, if the Base date/time is selected within 24 hours after the post was published, then *CC2* equals to *CC4*, shown in Figure 1. From Figure 1, we can observe that most of the comment are posted with the first 24 hour after the post was published. And it meets the assumption of "Only comments posted within the last three days relative to a given base date/time are relevant, as older posts are not typically expected to gain further engagement."

Feature Engineering:

The analysis aimed to tackle with a count data problem. In a Poisson regression model, offset is often considered when designing the regression. After observing the *Target_Variable* to the *Base_Time,* shown in Figure 2. Refer to Appendix.2, Base_Time means the hour from the selected Base date/time and to the time of the post published. We can see that the *Target_Variable* looks like a cumulative volume estimation of the post comment, so an offset term is considered in this analysis.

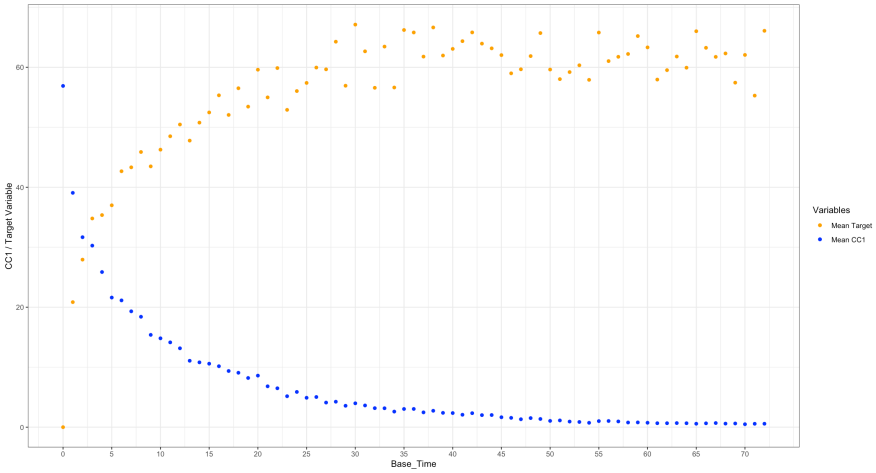


Figure 2. Patterns of Mean Target_Variable and CC1 in each Base_Time

Being interested in the average comment volume within every hour w.r.t the selected Base date/time, three new variables are created, *CC2_per_hr, CC3_per_hr, CC4_per_hr*, shown in Figure 3.

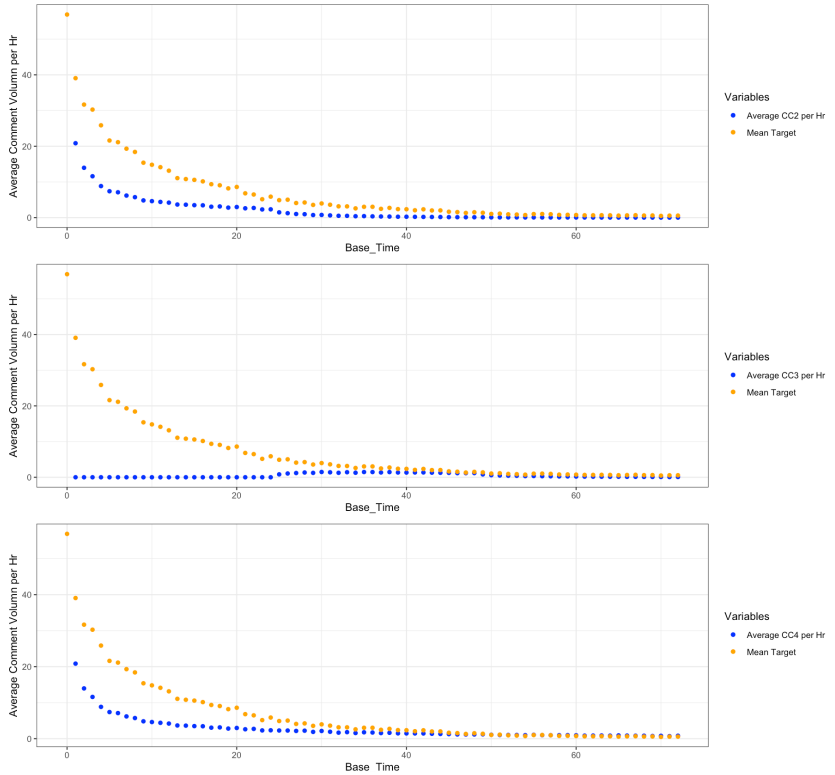


Figure 3. CC2_per_hr, CC3_per_hr, CC4_per_hr to Base_Time

Scales of predictors in regression analysis are also critical to estimate the estimands. The original data are highly skewed, and scale (range of the column) varies dramatically between variables. From Figure 4, I projected the data points to *CC1* and *Page_Popularity_Likes*. It is

Figure 4. Data projection before log transformation and normalization

obvious that the data scale varies drastically, and high skewness occurs. So I decided to perform a log transformation, and normalize the variable to make the distribution looks more like a bell-shaped, shown in Figure 5. *CC1, CC2, CC3, CC4, Page_Popularity_Likes, Page_Checkins, Page_Talking_About, Post_Length, Post_Share_Count, CC2_per_hr, CC3_per_hr, CC4_per_hr* are addressed by the log transformation and normalization. From Figure 5, we can see that the serious skewness and largely varying scale of the axis range had been mitigated.

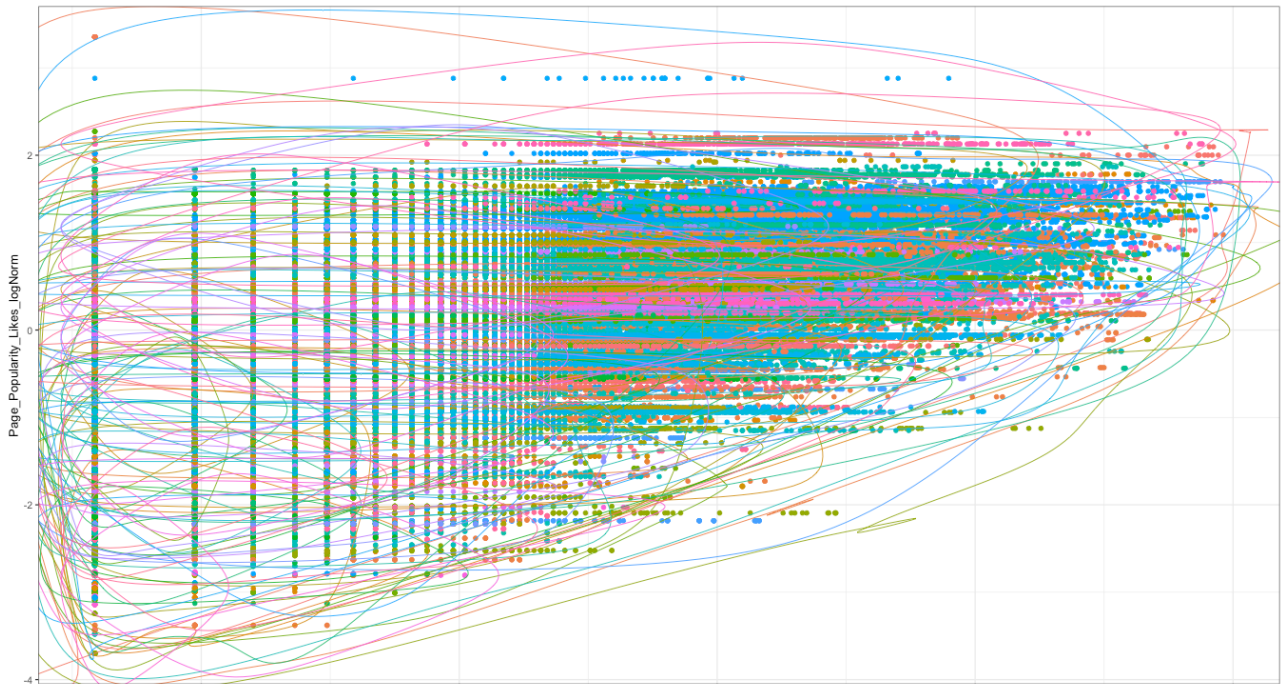


Figure 5. Data projection after log transformation and normalization

Regression Analysis:

After creating new variables and solved the data scaling problems, I started the regression analysis. Considering the meaning of the variables, the final predictors are listed below.

- CC1_logNorm
- CC2_logNorm
- CC3_logNorm
- CC4_logNorm
- CC5
- Page_Popularity_Likes_logNorm
- Page_Checkins_logNorm
- Page_Talking_About_logNorm
- Post_Length_logNorm
- Post_Share_Count_logNorm
- CC2_per_hr_logNorm
- CC3_per_hr_logNorm
- CC4_per_hr_logNorm

And by observing different mean *Target_Variable* from the data, shown in Figure 6. A mixed effect model is planned to adopt in the analysis, and the random effect of the *Page_Category* will be estimated in the intercept to present the randomness of each *Category*.

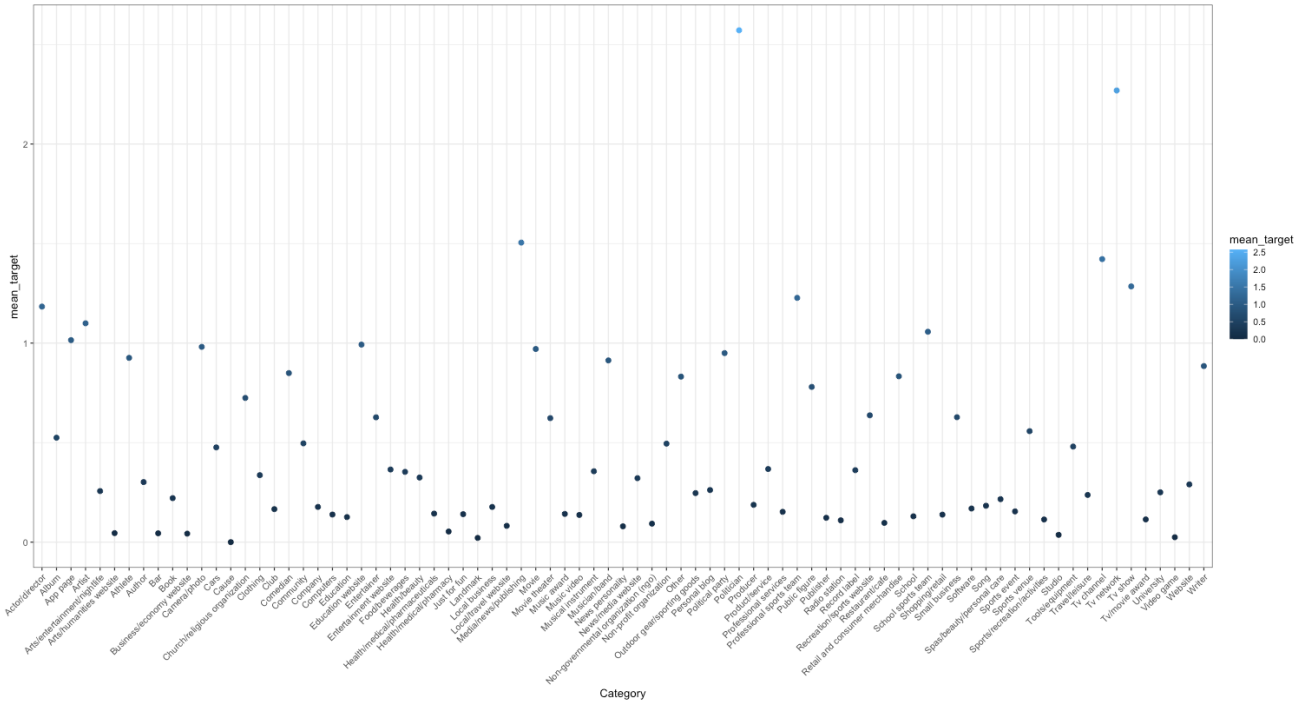


Figure 6. Mean Target_Variable of each Page_Category

Initially, a Poisson regression, a classical approach for count data analysis, was employed. However, issues related to varying scales persisted in the data, rendering complete removal infeasible, leading to the Poisson model's inability to converge for coefficient estimation. Furthermore, the presence of overdispersion contradicted the fundamental Poisson assumption where the parameter λ equals both the mean and the variance $\mu = var(y)$. Contrastingly, the observed mean and variance of the response variable, *Target_Variable*, were 7.17 and 1176.37, respectively. Consequently, both Quasi-Poisson regression and negative binomial regression were implemented to tackle the

overdispersion challenge. Additionally, a significant zero-inflation issue was identified during the fitting of the negative binomial model, which hindered estimation due to an excessive number of zeros. To address this, both a mixed-effect zero-inflation negative binomial regression model and a zero-inflation negative binomial regression model were utilized for the estimation process.

In summary, while two models, the Poisson regression and negative binomial regression, failed to complete fitting successfully, three others—Quasi-Poisson regression, zero-inflation negative binomial regression, and mixed-effect zero-inflation negative binomial regression model—achieved successful estimations.

Results:

Mixed effect zero-inflation negative binomial regression:

Table 1 and table 2 shows the result of estimation from the mixed effect zero-inflation negative binomial regression.

Table 1. Mixed Effect Zero Inflation Negative Binomial Regression Estimation - Counting Model

	Estimate	**Std.Error**	**z-value**	**Pr(>|z|)**	**Sig. Codes**[3]
(Intercept)	-2.0325	9.55E-02	-21.278946	1.78E-100	***
CC1_logNorm	2.1273	1.10E-01	19.333822	2.79E-83	***
CC2_logNorm	1.3555	1.58E-02	85.547878	0.00E+00	***
CC3_logNorm	-0.3662	1.67E-02	-21.951131	8.45E-107	***
CC4_logNorm	-6.6864	1.09E-01	-61.130051	0.00E+00	***
CC5	0.0023	5.46E-05	42.490858	0.00E+00	***
Page_Popularity_Likes_logNorm	0.1640	9.88E-03	16.596496	7.39E-62	***
Page_Checkins_logNorm	-0.0657	5.08E-03	-12.942558	2.59E-38	***
Page_Talking_About_logNorm	0.4274	1.03E-02	41.564859	0.00E+00	***
Post_Length_logNorm	-0.0096	4.68E-03	-2.046945	4.07E-02	*
Post_Share_Count_logNorm	0.5317	5.10E-03	104.155868	0.00E+00	***
CC2_per_hr_logNorm	-1.3008	2.21E-02	-58.77629	0.00E+00	***
CC3_per_hr_logNorm	0.3840	1.36E-02	28.299609	3.49E-176	***
CC4_per_hr_logNorm	3.5057	2.42E-02	144.826344	0.00E+00	***

Table 2. Mixed Effect Zero Inflation Negative Binomial Regression Estimation - Zero Inflation Model

	Estimate	**Std.Error**	**z-value**	**Pr(>|z|)**	**Sig. Codes**
(Intercept)	-6.2229	1.30E-01	-47.90495	0.00E+00	***

[3] Sig. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

CC1_logNorm	-0.2641	1.26E-01	-2.101095	3.56E-02	*
CC2_logNorm	-0.8902	6.57E-02	-13.545126	8.47E-42	***
CC3_logNorm	-0.1929	8.00E-02	-2.409282	1.60E-02	*
CC4_logNorm	2.3936	1.38E-01	17.337545	2.45E-67	***
CC5	0.0257	1.60E-03	16.102157	2.46E-58	***
Page_Popularity_Likes_logNorm	-0.0243	1.62E-02	-1.50385	1.33E-01	
Page_Checkins_logNorm	0.0297	1.07E-02	2.774515	5.53E-03	**
Page_Talking_About_logNorm	-0.2316	1.71E-02	-13.513238	1.31E-41	***
Post_Length_logNorm	-0.0739	9.23E-03	-8.004249	1.20E-15	***
Post_Share_Count_logNorm	-0.4986	1.51E-02	-33.044689	1.85E-239	***
CC2_per_hr_logNorm	1.8643	3.63E-01	5.133092	2.85E-07	***
CC3_per_hr_logNorm	0.7978	2.57E-01	3.10672	1.89E-03	**
CC4_per_hr_logNorm	-13.8657	3.77E-01	-36.774507	4.72E-296	***

The coefficient contains two parts of the estimation, one for counting model, which is the negative binomial regression model, and the second part is the estimation for extra zeros, which is the zero-inflation model. In prediction, for a given observation, the overall prediction involves using both parts of the model:

- If the zero-inflation model predicts a high probability for a zero (π_i is high), the observation is likely to be one of the extra zeros.
- If the zero-inflation model predicts a low probability for a zero (π_i is low), then the count prediction from the Negative Binomial part (μ_i) is used.

By focusing on the coefficients of *Page_Popularity_Likes_logNorm, Page_Talking_About_logNorm*, and *Post_Share_Count_logNorm*, it is evident that their magnitudes are significant enough to positively affect the response variable. Additionally, their small p-values underscore their statistical significance. Conversely, in the zero-inflation model, the coefficient of *CC4_per_hr_logNorm* is worth noting. This coefficient is remarkably large and is supported by a very small p-value, indicating its significance. This suggests that an increase in comment volume within the first 24 hours after a post's publication is associated with a lower likelihood of predicting an extra zero for that instance. Random effects for each *Page_Category* is shown in Appendix 4.

Zero-inflation negative binomial regression:

It is interesting to validate the result whether the *Page_Category* truly cause random effects among the data. So, a zero-inflation negative binomial regression is also fitted. Table 3 and Table 4 shows the result of estimation from the zero-inflation negative binomial regression.

Table 3. Zero Inflation Negative Binomial Regression Estimation - Counting Model

	Estimate	Std.Error	z-value	Pr(>|z|)	Sig. Codes
(Intercept)	-1.5592	6.15E-03	-253.45914	0.00E+00	***
CC1_logNorm	2.2077	1.15E-01	19.254813	1.29E-82	***
CC2_logNorm	1.4049	1.63E-02	86.227307	0.00E+00	***
CC3_logNorm	-0.3138	1.73E-02	-18.168251	9.21E-74	***
CC4_logNorm	-6.7877	1.14E-01	-59.666976	0.00E+00	***
CC5	0.0023	5.47E-05	41.35275	0.00E+00	***
Page_Popularity_Likes_logNorm	0.0244	7.82E-03	3.117069	1.83E-03	**
Page_Checkins_logNorm	-0.1414	4.19E-03	-33.733925	1.84E-249	***
Page_Talking_About_logNorm	0.4739	7.97E-03	59.479024	0.00E+00	***
Post_Length_logNorm	-0.0312	4.61E-03	-6.778669	1.21E-11	***
Post_Share_Count_logNorm	0.5100	4.87E-03	104.694198	0.00E+00	***
CC2_per_hr_logNorm	-1.3399	2.27E-02	-59.026349	0.00E+00	***
CC3_per_hr_logNorm	0.3398	1.40E-02	24.284428	2.86E-130	***
CC4_per_hr_logNorm	3.5631	2.48E-02	143.568011	0.00E+00	***
Log(theta)[4]	-0.3333	5.39E-03	-61.869171	0.00E+00	***

Table 4. Zero Inflation Negative Binomial Regression Estimation – Zero Inflation Model

	Estimate	Std.Error	z-value	Pr(>|z|)	Sig. Codes
(Intercept)	-7.2710	9.68E-02	-75.148371	0.00E+00	***
CC1_logNorm	-2.4026	1.29E-01	-18.554244	7.54E-77	***
CC2_logNorm	-0.5172	6.17E-02	-8.3790124	5.34E-17	***
CC3_logNorm	-0.0153	7.46E-02	-0.2048657	8.38E-01	
CC4_logNorm	2.2770	1.39E-01	16.3308768	5.95E-60	***
CC5	0.0157	1.42E-03	11.0352273	2.58E-28	***
Page_Popularity_Likes_logNorm	-0.0374	1.71E-02	-2.1853984	2.89E-02	*
Page_Checkins_logNorm	0.0060	1.13E-02	0.5289232	5.97E-01	
Page_Talking_About_logNorm	-0.0655	1.80E-02	-3.6337939	2.79E-04	***
Post_Length_logNorm	-0.0410	9.78E-03	-4.1915444	2.77E-05	***
Post_Share_Count_logNorm	-0.2595	1.57E-02	-16.532687	2.13E-61	***
CC2_per_hr_logNorm	1.6518	3.30E-01	5.0061551	5.55E-07	***
CC3_per_hr_logNorm	0.3787	2.28E-01	1.6591088	9.71E-02	.
CC4_per_hr_logNorm	-8.3536	3.41E-01	-24.487301	2.02E-132	***

[4] θ is known as a dispersion parameter in GLM.

In the zero-inflation negative binomial regression model, coefficients for *Page_Talking_About_logNorm* and *Post_Share_Count_logNorm* positively influence the increase in comment volume. However, an intriguing observation emerges when examining the coefficients of *CC4_logNorm* and *CC4_per_hr_logNorm*. These coefficients display contrasting effects. Specifically, the model predicts a positive correlation between *CC4_logNorm* and the *Target_Variable*, which represents the total number of comments within the first 24 hours. In contrast, there is a negative correlation between *CC4_per_hr_logNorm* and the *Target_Variable*, which represent the comment volume in each hour on average. This finding presents a conflict with the predictions from the Negative Binomial model, which includes the random variability of *Page_Category*.

Quasi-Poisson Regression:

Despite the prevalence of zeros in the data, I initially overlooked them at the beginning of the study. Therefore, a quasi-Poisson regression somewhat reflects my preliminary assumptions about the data distribution and the design of the regression analysis.

Table 5. Quasi-Poisson Regression Estimation

	Estimate	**Std.Error**	**z-value**	**Pr(>|z|)**	**Sig. Codes**
(Intercept)	-2.9999	4.29E-02	-69.960586	0.00E+00	***
CC1_logNorm	-2.6908	2.95E-01	-9.133883	6.67E-20	***
CC2_logNorm	2.4835	1.18E-01	21.050483	2.90E-98	***
CC3_logNorm	0.3500	6.64E-02	5.271062	1.36E-07	***
CC4_logNorm	-1.2094	2.79E-01	-4.341512	1.42E-05	***
CC5	-0.0008	7.04E-05	-11.557283	6.94E-31	***
Page_Popularity_Likes_logNorm	-0.0821	3.22E-02	-2.550686	1.08E-02	*
Page_Checkins_logNorm	-0.1706	1.43E-02	-11.965933	5.50E-33	***
Page_Talking_About_logNorm	0.8067	4.05E-02	19.915721	3.63E-88	***
Post_Length_logNorm	0.0253	1.67E-02	1.510022	1.31E-01	
Post_Share_Count_logNorm	0.4507	1.85E-02	24.412084	1.99E-131	***
CC2_per_hr_logNorm	-0.8208	1.08E-01	-7.619805	2.55E-14	***
CC3_per_hr_logNorm	-0.3316	4.41E-02	-7.511793	5.86E-14	***
CC4_per_hr_logNorm	2.3337	1.16E-01	20.122124	5.81E-90	***

In the Quasi-Poisson regression, coefficients for *Page_Talking_About_logNorm* and *Post_Share_Count_logNorm* are substantial and positively impact the *Target_Variable*. Furthermore, the coefficient for *CC4_per_hr_logNorm*, representing the average CC4 per hour within the first 24 hours, exhibits a positive effect. This contrasts with the coefficient for *CC4_logNorm*, which influences the *Target_Variable* in a negative direction.

Feature Selection:

After interpreting the results from these three models, it became evident that there is a significant correlation among the essential features (*CC1*, *CC2*, *CC3*, *CC4*), contributing to the volatility in estimation. Consequently, upon reviewing the correlation matrix for the selected covariates, as presented in Appendix 5, I decided to eliminate *CC1*, *CC2*, *CC3*, and their associated variables. The refined list of variables is as follows:

- CC4_logNorm
- CC5
- Page_Popularity_Likes_logNorm
- Page_Checkins_logNorm
- Page_Talking_About_logNorm
- Post_Length_logNorm
- Post_Share_Count_logNorm
- CC4_per_hr_logNorm

Accordingly, a mixed-effect zero-inflation negative binomial model will be applied to analyze this new set of covariates. Table 6 and Table 7 show the estimation result of the Feature Selected Mixed Effect Zero-Inflation Negative Binomial regression.

Table 6. Feature Selected Mixed Effect Zero Inflation Negative Binomial Regression Model – Condition Model

	Estimate	**Std.Error**	**z-value**	**Pr(>|z|)**	**Sig. Codes**
(Intercept)	-0.7815	1.18E-01	-6.614885	3.72E-11	***
CC4_logNorm	-0.7090	6.61E-03	-107.25271	0.00E+00	***
CC5	0.0036	2.92E-05	124.605116	0.00E+00	***
Page_Popularity_Likes_logNorm	0.4098	1.72E-02	23.810115	2.62E-125	***
Page_Checkins_logNorm	-0.1105	9.26E-03	-11.937018	7.59E-33	***
Page_Talking_About_logNorm	0.4901	1.66E-02	29.595444	1.71E-192	***
Post_Length_logNorm	-0.0447	8.72E-03	-5.125318	2.97E-07	***
Post_Share_Count_logNorm	0.8150	8.73E-03	93.314913	0.00E+00	***

Table 7. Feature Selected Mixed Effect Zero Inflation Negative Binomial Regression Model – Zero Inflation Model

	Estimate	**Std.Error**	**z-value**	**Pr(>|z|)**	**Sig. Codes**
(Intercept)	-0.7815	1.18E-01	-6.614885	3.72E-11	***
CC4_logNorm	-0.7090	6.61E-03	-107.25271	0.00E+00	***
CC5	0.0036	2.92E-05	124.605116	0.00E+00	***
Page_Popularity_Likes_logNorm	0.4098	1.72E-02	23.810115	2.62E-125	***
Page_Checkins_logNorm	-0.1105	9.26E-03	-11.937018	7.59E-33	***
Page_Talking_About_logNorm	0.4901	1.66E-02	29.595444	1.71E-192	***
Post_Length_logNorm	-0.0447	8.72E-03	-5.125318	2.97E-07	***
Post_Share_Count_logNorm	0.8150	8.73E-03	93.314913	0.00E+00	***

Focusing on *CC4_logNorm*, the coefficient is observed to be negative. This can be attributed to the relationship between the number of comments received within the first 24 hours after a post is published and a fixed audience size. It implies that an increase in comment volume within this initial 24-hour period leads to a diminished effect on the overall hourly comment volume. This interpretation takes into consideration the offset term included in the model, suggesting that a higher initial comment volume negatively impacts subsequent hourly comment activity.

A Bayesian based mixed effect negative binomial regression is also fitted for comparison among different evaluation metrics.

Evaluation Metrics:

Table 8 shows the four performance metrics by the fitted data of regression models for each model.

Table 8. Performance Metric of Fitted Data

Model	MSE	MAPE	RMSE	MAE
Mixed Effect Zero Inflation Negative Binomial	2,180,620	6.63	1,476.69	25.17
Zero Inflation Negative Binomial	11,461,908	9.40	3,385.54	33.51
Quasi-Poisson	783.83	0.93	28.00	5.46
Feature Selected Mixed Effect Zero Inflation Negative Binomial	492,908	22.94	702.07	45.73
Null Model - Negative Binomial[5]	1,176.39	4.34	34.30	10.66
Negative Binomial - Bayesian Approach	1,255.77	0.11	35.44	5.14

It is easy to observe that the Mixed Effect Zero Inflation Negative Binomial, Zero Inflation Negative Binomial, and Feature Selected Mixed Effect Zero Inflation Negative Binomial models exhibit large Mean Squared Errors (MSE). This is primarily due to the extremely large fitted values, which result in substantial residuals after squaring, an issue that also affects the Root Mean Squared Error (RMSE). Interestingly, all these models incorporate Zero Inflation model estimation, which, in this case, appears to adversely impact model fitting. The challenge seems to arise from the models' inability to simultaneously capture the extra zeros and the long-tailed distribution of the response variable. Focusing more on the extra zeros seems to actually generate more errors, highlighting a limitation in these models' capacity to balance these aspects effectively.

Residual Analysis:

In Generalized Linear Models (GLMs), the classical assumptions of normality and independence and identically distributed (iid) residuals, fundamental to ordinary linear regression, are modified.

[5] Null Model Summary: Please refer to Appendix 6.

GLMs accommodate response variables following various distributions from the exponential family, such as binomial, Poisson, and normal. This framework incorporates a link function, connecting the mean of the response variable to the linear predictor. Crucially, the assumption of normally distributed residuals is relaxed; instead, residuals are expected to exhibit a distribution consistent with the chosen response distribution. While the normality of residuals is not a prerequisite, the independence of observations remains a critical assumption. Additionally, GLMs often account for heteroscedasticity, where the variance of residuals is a function of the mean, in contrast to the homoscedasticity assumption in linear regression. Residuals plots of each model are attached in Appendix 7 to Appendix 11.

Prediction:

Evaluation Metrics:

Table 9 shows the predictive result of the four performance metric for each model.

Table 9. Prediction Performance Metrics

Model	MSE	MAPE	RMSE	MAE
Mixed Effect Zero Inflation Negative Binomial	11,620.43	3.68	107.80	29.77
Zero Inflation Negative Binomial	11,573.89	3.62	107.58	29.47
Quasi-Poisson	11,656.87	0.09	107.97	26.69
Feature Selected Mixed Effect Zero Inflation Negative Binomial	13,421,180	9.43	3,663.49	252.17
Null Model - Negative Binomial[6]	11,347.33	2.81	106.52	27.35
Negative Binomial - Bayesian Approach	11,660.67	0.47	107.98	26.51

The predictive analysis demonstrated that the Quasi-Poisson model, excluding Bayesian regression, outperformed all others in terms of Mean Absolute Percentage Error (MAPE). Interestingly, even though zero inflation issues were identified and addressed by specialized models, the inclusion of a zero inflation model did not markedly improve model fit in this instance. In fact, it makes things worse. Furthermore, it was observed that removing variables from the model did not enhance its performance compared to the original setup. This finding contradicts the initial thought of overfitting, as the performance metrics consistently reflected similar predictive results. Additionally, it is unfortunate that weekday features were excluded due to ambiguous definitions, a decision that potentially limited the interpretability of the findings.

[6] Null Model Summary: Please refer to Appendix 6.

Conclusion:

In conclusion, this research provides a comprehensive analysis of various regression models for predicting Facebook comment volumes. The study reveals the effectiveness of the Quasi-Poisson model in handling data with zero inflation issues, as evidenced by its superior performance in terms of Mean Absolute Percentage Error (MAPE). Additionally, the exploration of the Mixed Effect Zero-Inflation Negative Binomial model and Bayesian-based approaches offers insights into their capabilities and limitations in addressing the complexities of the data. The research underscores the effect of feature selection, as demonstrated by the decision to eliminate certain variables, which significantly impacted model outcomes. The findings highlight the nuanced balance needed in model selection and feature inclusion, especially when dealing with social media data characterized by zero inflation and long-tailed distributions. This study provides practical implications for effectively predicting engagement metrics like comment volumes on platforms like Facebook using regressions.

References:

- Singh, Kamaljot, Ranjeet Kaur Sandhu, and Dinesh Kumar. "Comment volume prediction using neural networks and decision trees." IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015 (UKSim2015). 2015.
- Singh, Kamaljot. "Facebook comment volume prediction." *International Journal of Simulation: Systems, Science and Technologies* 16.5 (2015): 16-1.

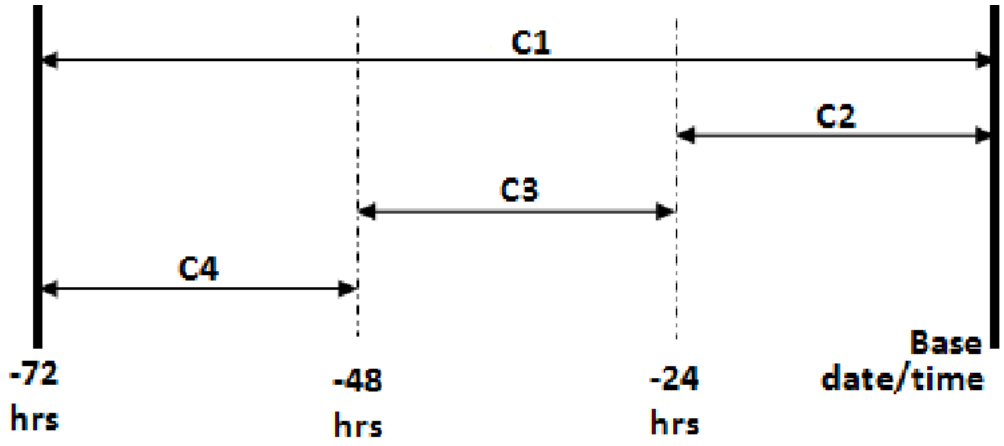
Appendix 1. Summary Statistics

Data Variant 5

Statistic	N	Mean	St.Dev.	Min	Max
Page_Category	199,030	24.242	19.935	1	106
Page_Popularity_Likes	199,030	1,313,785.00	6,771,131.00	36	486,972,297
Page_Checkins	199,030	4,674.52	20,573.44	0	186,370
Page_Talking_About	199,030	44,771.73	110,898.30	0	6,089,942
CC1_Min	199,030	0.47	13.178	0	1,458
CC1_Max	199,030	485.318	538.194	0	2,495
CC1_Avg	199,030	55.901	86.515	0	2,031.00
CC1_Median	199,030	35.264	68.163	0	2,123.00
CC1_Std	199,030	68.091	82.411	0	762.358
CC2_Min	199,030	0.068	2.173	0	227
CC2_Max	199,030	381.499	439.634	0	2,119
CC2_Avg	199,030	21.815	35.693	0	973.25
CC2_Median	199,030	7.17	19.701	0	1,121.00
CC2_Std	199,030	40.514	51.561	0	683.596
CC3_Min	199,030	0.006	0.872	0	148
CC3_Max	199,030	380.723	430.183	0	2,095
CC3_Avg	199,030	19.992	31.568	0	660.75
CC3_Median	199,030	4.876	13.072	0	487
CC3_Std	199,030	40.712	52.598	0	801.468
CC4_Min	199,030	0.469	13.126	0	1,458
CC4_Max	199,030	434.882	490.73	0	2,184
CC4_Avg	199,030	52.754	81.02	0	1,868.50
CC4_Median	199,030	33.608	64.178	0	1,992.50
CC4_Std	199,030	63.461	76.836	0	680.962
CC5_Min	199,030	-326.275	380.145	-2,038	0
CC5_Max	199,030	377.323	436.702	-101	2,119
CC5_Avg	199,030	1.822	9.69	-184.4	496.6
CC5_Median	199,030	-2.119	10.488	-175	521
CC5_Std	199,030	56.54	74.583	0	1,386.40
CC1	199,030	55.901	137.524	0	2,495
CC2	199,030	21.815	74.658	0	2,119
CC3	199,030	19.992	73.625	0	2,095
CC4	199,030	52.754	128.434	0	2,184

CC5	199,030	1.822	94.092	-2,038	2,119
Base_Time	199,030	35.45	21.006	0	72
Post_Length	199,030	163.692	375.663	0	21,480
Post_Share_Count	199,030	117.363	954.359	1	144,860
Post_Promotion_Status	199,030	0	0	0	0
H_Local	199,030	23.783	1.827	1	24
Post_Published_Weekday_40	199,030	0.122	0.328	0	1
Post_Published_Weekday_41	199,030	0.143	0.35	0	1
Post_Published_Weekday_42	199,030	0.149	0.357	0	1
Post_Published_Weekday_43	199,030	0.157	0.364	0	1
Post_Published_Weekday_44	199,030	0.144	0.351	0	1
Post_Published_Weekday_45	199,030	0.146	0.353	0	1
Post_Published_Weekday_46	199,030	0.137	0.344	0	1
Base_DateTime_Weekday_47	199,030	0.139	0.346	0	1
Base_DateTime_Weekday_48	199,030	0.135	0.342	0	1
Base_DateTime_Weekday_49	199,030	0.137	0.344	0	1
Base_DateTime_Weekday_50	199,030	0.147	0.354	0	1
Base_DateTime_Weekday_51	199,030	0.155	0.362	0	1
Base_DateTime_Weekday_52	199,030	0.144	0.351	0	1
Base_DateTime_Weekday_53	199,030	0.142	0.349	0	1
Target_Variable	199,030	7.169	34.298	0	1,702
CC2_per_hr	199,030	2.098	10.597	0	1,011.00
CC3_per_hr	199,030	0.504	1.92	0	58.6
CC4_per_hr	199,030	2.768	10.785	0	1,011.00
CC1_logNorm	199,030	0	1	-1.413	2.962
CC2_logNorm	199,030	0	1	-0.929	3.866
CC3_logNorm	199,030	0	1	-0.767	3.943
CC4_logNorm	199,030	0	1	-1.397	2.93
Page_Popularity_Likes_logNorm	199,030	0	1	-3.699	3.352
Page_Checkins_logNorm	199,030	0	1	-0.643	2.838
Page_Talking_About_logNorm	199,030	0	1	-2.717	2.364
Post_Length_logNorm	199,030	0	1	-2.345	3.323
Post_Share_Count_logNorm	199,030	0	1	-1.112	4.901
CC2_per_hr_logNorm	199,030	0	1	-0.542	7.999
CC3_per_hr_logNorm	199,030	0	1	-0.465	8.335
CC4_per_hr_logNorm	199,030	0	1	-0.778	7.195

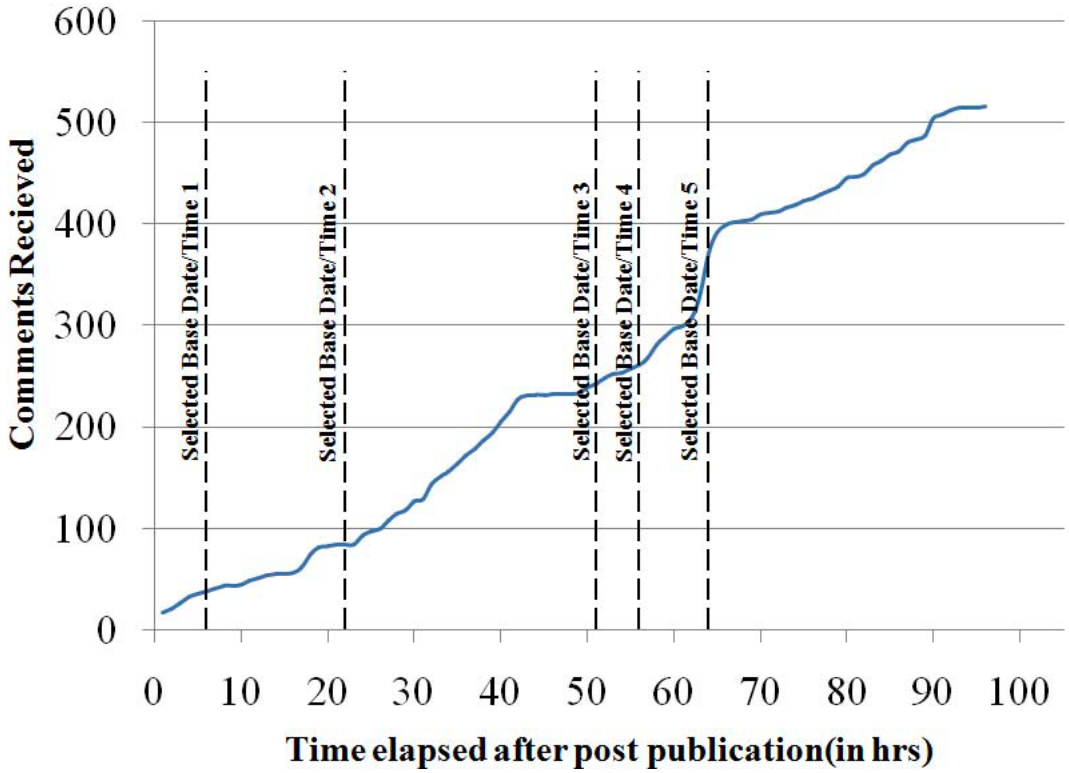
Appendix 2. Definition of CC1, CC2, CC3, CC4



Reference:

- Singh, Kamaljot, Ranjeet Kaur Sandhu, and Dinesh Kumar. "Comment volume prediction using neural networks and decision trees." IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015 (UKSim2015). 2015. Figure 2. Demonstrating the essential feature details.

Appendix 3. Selected Base date/time and Data Variants



Reference:

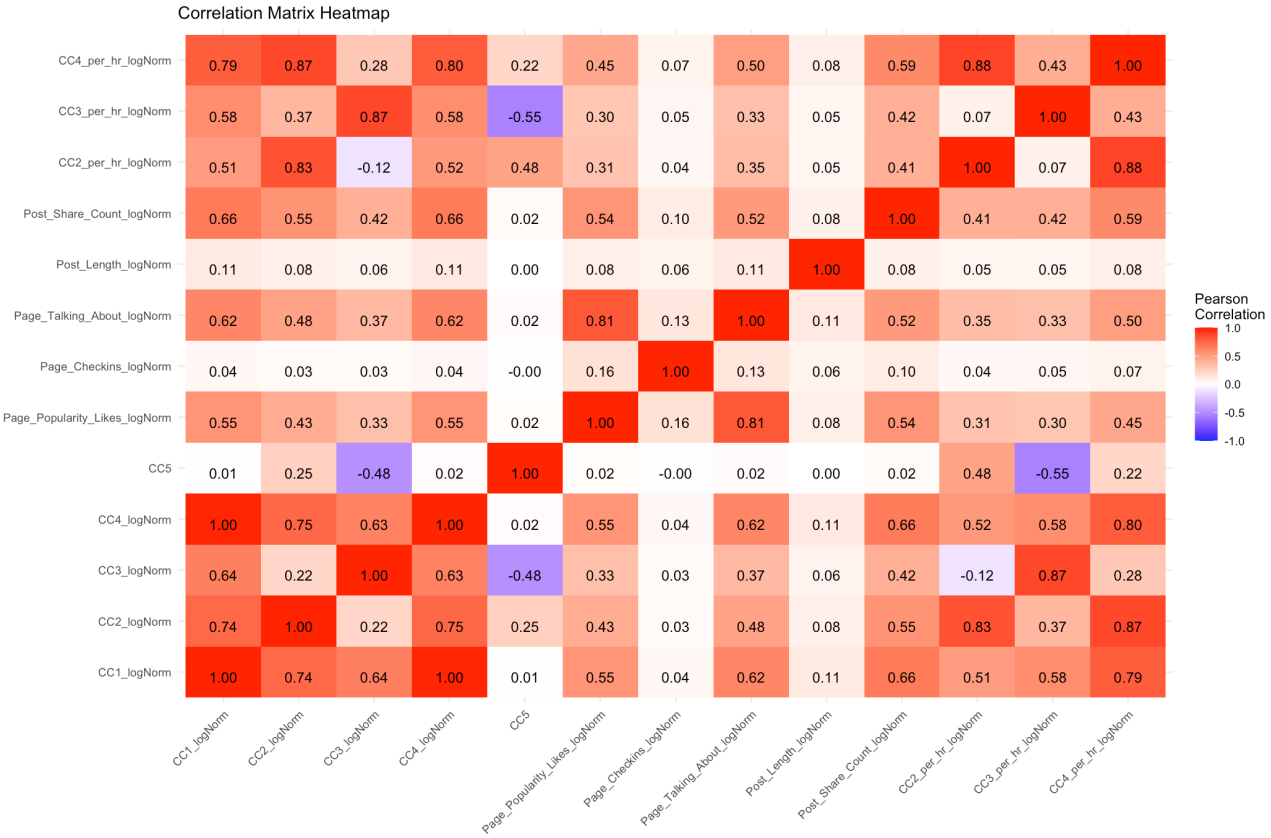
- Singh, Kamaljot, Ranjeet Kaur Sandhu, and Dinesh Kumar. "Comment volume prediction using neural networks and decision trees." IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015 (UKSim2015). 2015. Figure 3. Cumulative Comments and different selected base date/time.

Appendix 4. Random effect in intercept of Mixed Effect Zero-Inflation Negative Binomial Regression Model

	(Intercept)
Actor/director	0.797691416002855
Album	1.45481231819863
App page	0.335323126318331
Artist	0.752756533780161
Arts/entertainment/nightlife	0.570816872757582
Arts/humanities website	-0.722270126194794
Athlete	0.366450479118773
Author	-0.12106639763402
Bar	-0.928223930349228
Book	0.234561586233673
Business/economy website	-2.28007195665994
Camera/photo	-0.17352534636411
Cars	-0.865934554135371
Cause	-0.0309500108743655
Church/religious organization	0.93568102899322
Clothing	-0.465299017798948
Club	0.471516759561997
Comedian	-0.271051385439421
Community	0.922300765894157
Company	0.371478903710236
Computers	-0.29465486402889
Education	0.251259557259128
Education website	0.307220727423639
Entertainer	0.748213395751029
Entertainment website	-0.0257047879916132
Food/beverages	-0.690744976361128
Health/beauty	-0.273928873397458
Health/medical/pharmaceuticals	-0.498513361764386
Health/medical/pharmacy	-0.488630012219016
Just for fun	-0.131296619001435
Landmark	-0.755549492915212
Local business	-0.315145403445216
Local/travel website	-1.42795642304809
Media/news/publishing	0.93149958473011
Movie	0.463823621182598
Movie theater	1.17718881789451
Music award	-0.499557791753261
Music video	-0.0982061808845753

Musical instrument	0.830272757655659
Musician/band	0.653116383577719
News personality	-1.05180790649609
News/media website	-0.290747542809893
Non-governmental organization (ngo)	-0.645099275334324
Non-profit organization	-0.0363552617027611
Other	1.03530774989589
Outdoor gear/sporting goods	-0.823000907462748
Personal blog	-0.299202369530877
Political party	0.488278087640249
Politician	1.003187209019
Producer	-0.432381453067569
Product/service	-0.368180005813749
Professional services	0.154554392372236
Professional sports team	0.10894069597587
Public figure	0.884559827920183
Publisher	-2.0980898561788
Radio station	-0.456240048286791
Record label	0.119206624796966
Recreation/sports website	-0.448078625179603
Restaurant/cafe	0.0301647819194763
Retail and consumer merchandise	0.5275719809678
School	-0.32949395046109
School sports team	0.200767446671532
Shopping/retail	-0.94736319911479
Small business	0.14074599393938
Software	-0.793933506742064
Song	-0.418155110560129
Spas/beauty/personal care	-0.396258737086515
Sports event	0.146998127646654
Sports venue	1.77809218188632
Sports/recreation/activities	1.06258542110691
Studio	-0.106930344201876
Tools/equipment	0.0269472353308642
Travel/leisure	-1.05836457879078
Tv channel	0.706075037648395
Tv network	-0.0943614323275968
Tv show	2.15279209093202
Tv/movie award	-0.793964542442658
University	0.449464358638077
Video game	-1.5435171473406
Website	0.00814223070591014
Writer	0.814248014629271

Appendix 5. Correlation for Selected Covariates

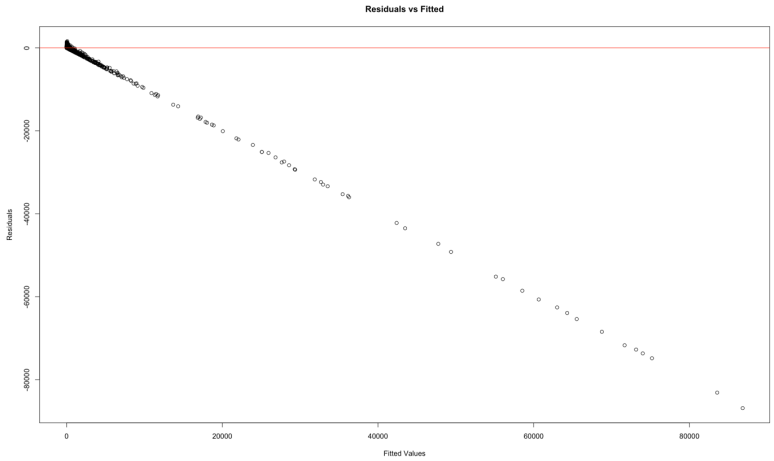


Appendix 6. Null Model Summary

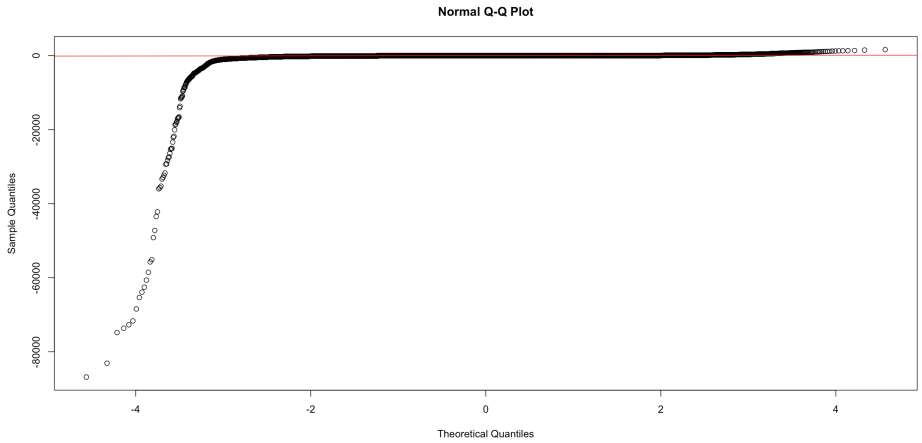
```
Family: nbinom2  ( log )
Formula:          Target_Variable ~ 1
Data: fv5_train

     AIC       BIC    logLik  deviance  df.resid
 845937.5  845957.9 -422966.7  845933.5    199028


Dispersion parameter for nbinom2 family (): 0.139

Conditional model:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.969725   0.006061     325   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 7. Residual Analysis - Feature Selected Mixed Effect Zero Inflation Negative Binomial Regression Model

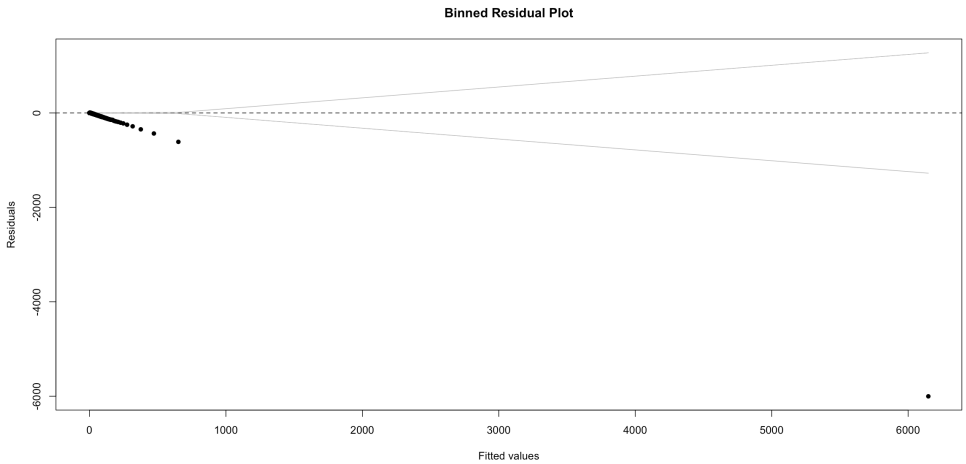
Residual Plots:



QQ-Plot:

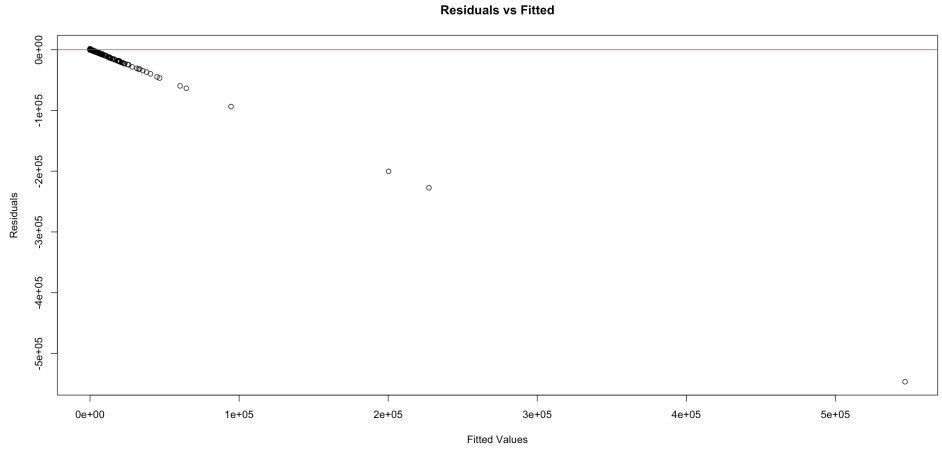


Binned Residual Plot:

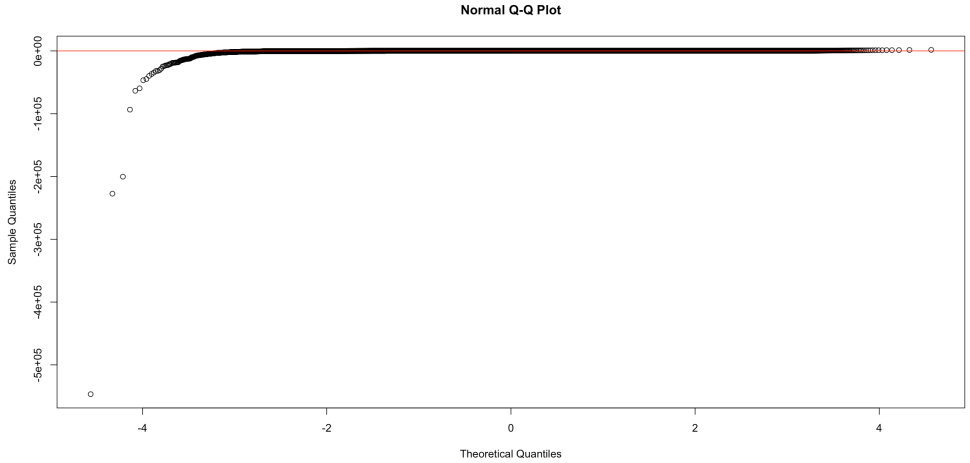


Appendix 8. Residual Analysis - Mixed Effect Zero Inflation Negative Binomial Regression Model

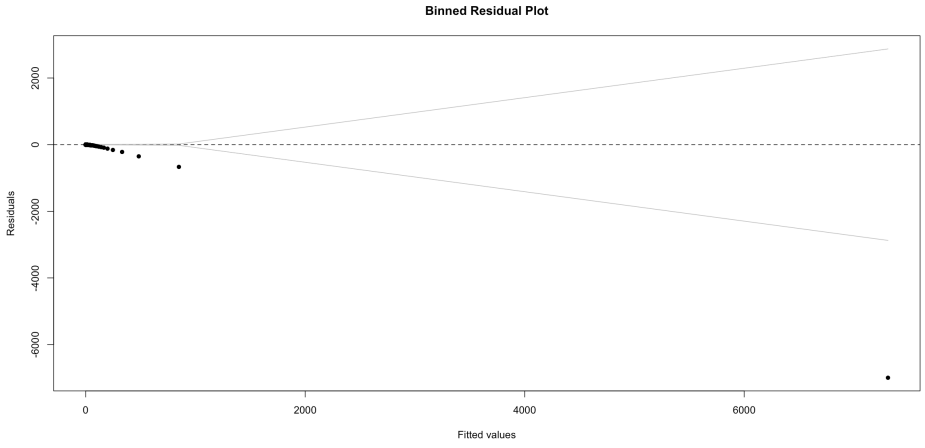
Residual Plots:



QQ-Plot:

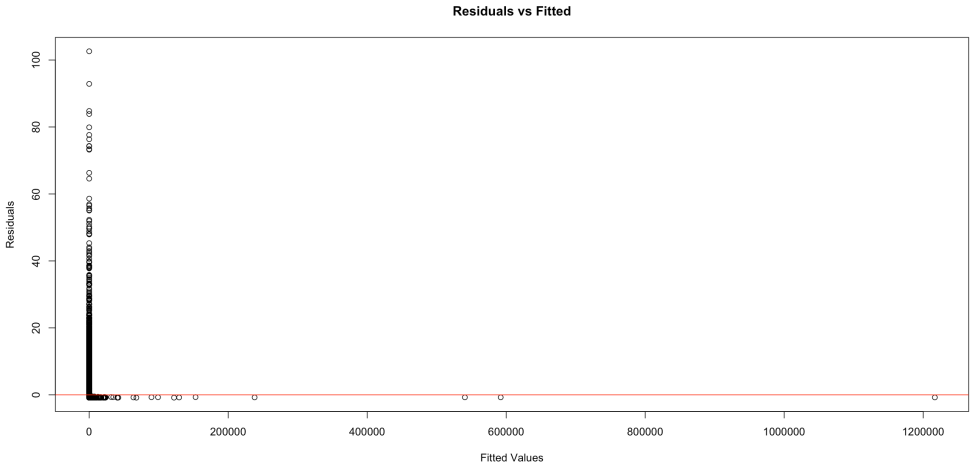


Binned Residual Plot:

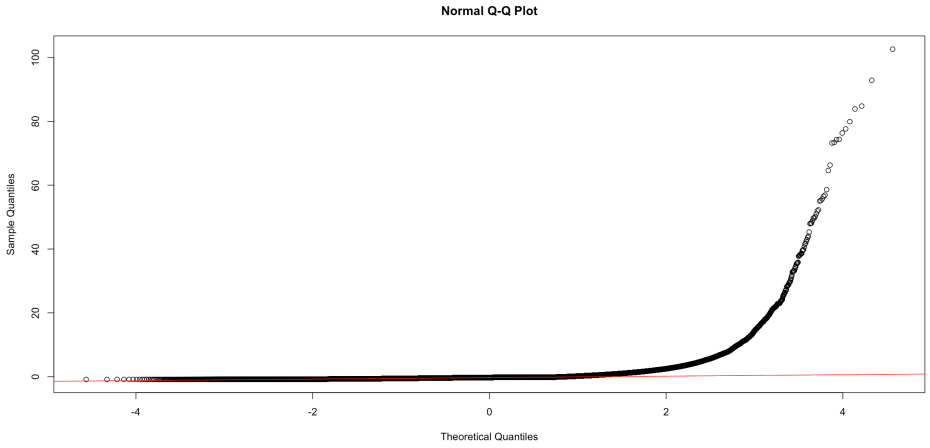


Appendix 9. Residual Analysis - Zero Inflation Negative Binomial Regression Model

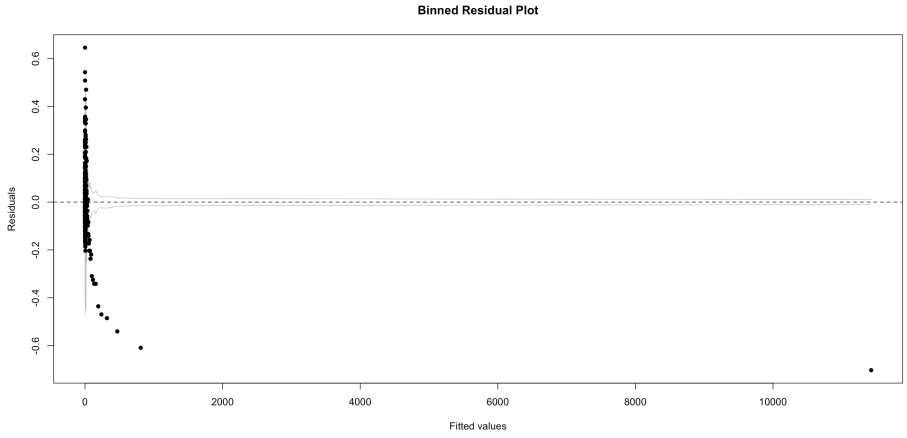
Residual Plots:



QQ-Plot:

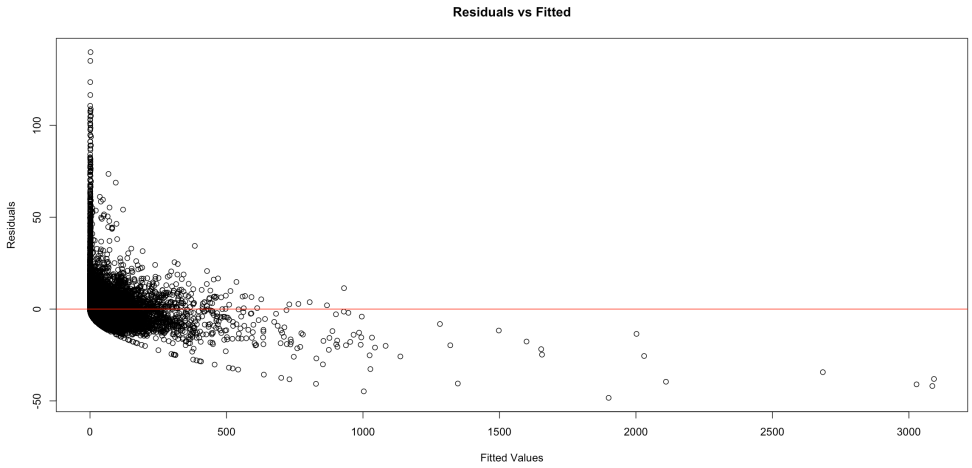


Binned Residual Plot:

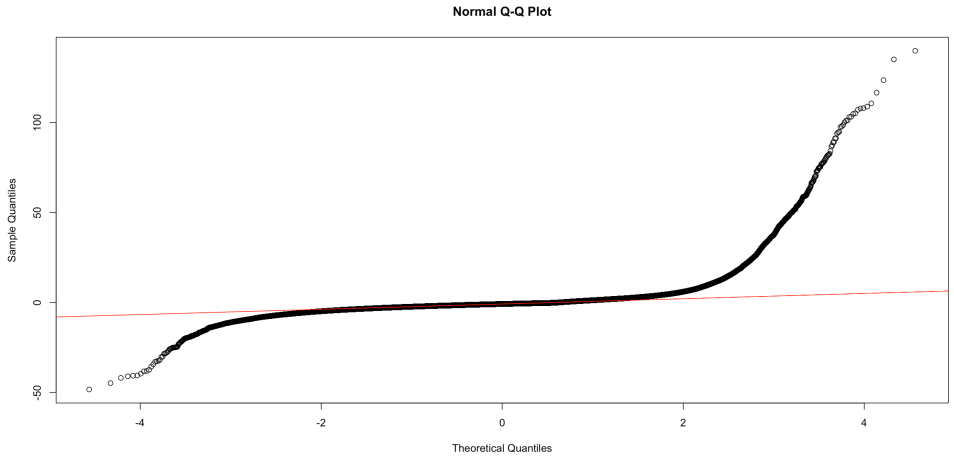


Appendix 10. Residual Analysis - Zero Inflation Negative Binomial Regression Model

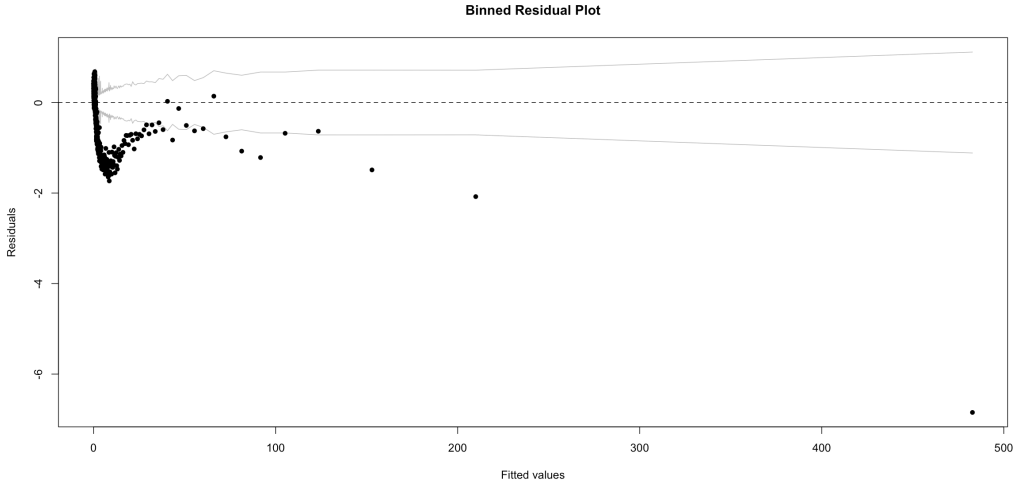
Residual Plots:



QQ-Plot:

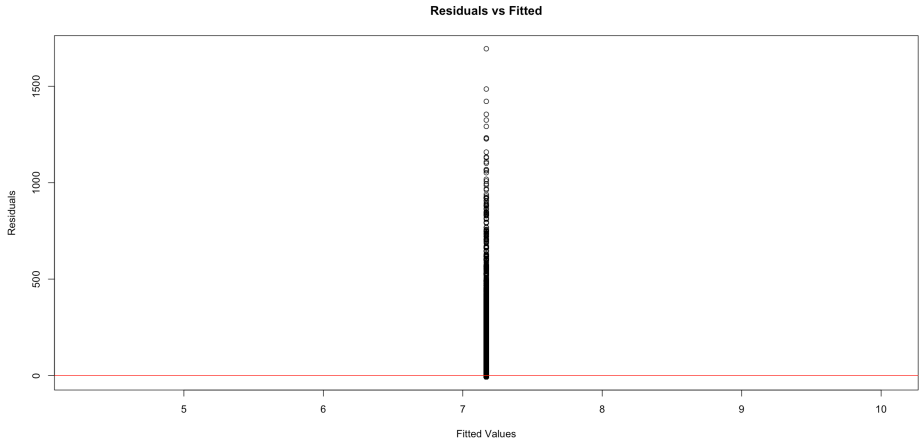


Binned Residual Plot:

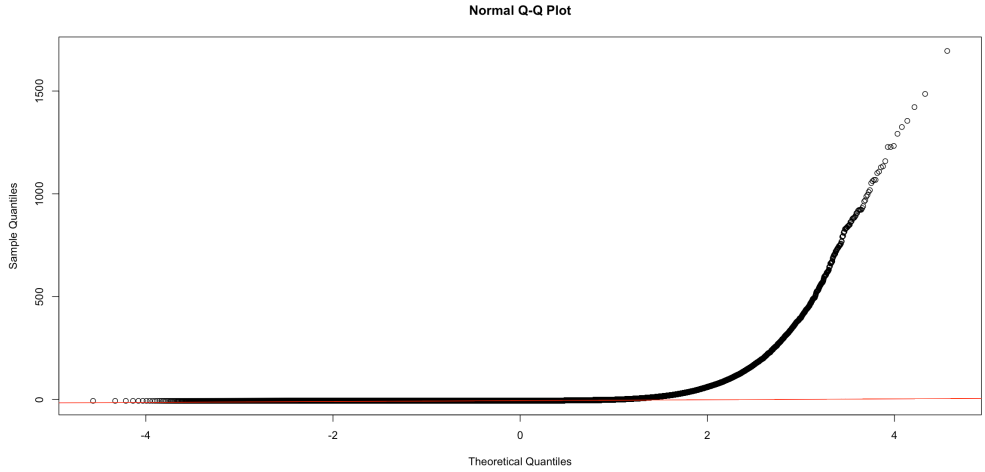


Appendix 11. Residual Analysis – Null Model (Negative Binomial Regression Model)

Residual Plots:



QQ-Plot:



Binned Residual Plot: