Facebook Post Comment Volume Regression Analysis

Ta-Hung (Denny) Chen

Department of Mathematics and Statistics

Boston University, Boston, Massachusetts, U.S

Instructor: Dr. AmirEmad Ghassami

December, 15, 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. Drawing on the work of Kamaljot Singh and others, my research employs a Bayesian approach to construct a count data regression model that predicts the number of comments a post is likely to receive within the subsequent hours of its publication, and remain flexibility to adjust the model 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 Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), alongside Posterior Predictive Check plots and Markov Chain Monte Carlo Diagnostics for Validation of the convergence and goodness of the fitting of the models. 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. Negative binomial regression is the methodology used due to an overdispersion observed in the data. 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.

Methodology:

Estimation:

In this research project, I initially employed Poisson regression to model count data. However, upon observing overdispersion, which indicate that the data mean is not equal to data variance. I shifted to a Negative Binomial regression model. This approach is more flexible for count data with overdispersion. In implementing Bayesian regression analysis for both Poisson and Negative Binomial models, I applied weakly informative priors. Specifically, I used a normal distribution with a large scale as the prior. This choice was intended to minimally influence the model outcomes while still providing enough structure to stabilize the estimates. To further refine the model, I allowed the rstan program to automatically adjust the scale of these priors. Moreover, for the Negative Binomial regression, I introduced an additional prior: the Laplace distribution to the fixed effect of the parameters. To the random effect, the prior of the covariance matrix will remain the default setting as applying LKJ distribution to the correlation matrix, and models the variances as the product of simplex vector (follows symmetric Dirichlet distribution where concentration set to 1 by default) and the trace of covariance matrix (product of the order of the covariance matrix and square of scale parameter follow $gamma(1,1)$ by default, which is a weakly informative prior). This was done to investigate the potential benefits of a different prior distribution on the regression model's performance.

The Laplace distribution is characterized by a peak at its median and exhibits exponential decay on either side of the peak, with the rate of decay being controlled by the scale parameter. The Laplace distribution is often used when a certain degree of sparsity is desired in the parameter estimates. This can be especially useful in high-dimensional settings where many parameters may be irrelevant and should ideally be shrunk towards zero.

Laplace:

$$f(\, x \mid \mu, b\,) = \frac{1}{2\lambda}\, exp(\frac{|x-\mu|}{-\lambda}\,)$$

where:

x is the variable

μ is the median of the distribution

λ is the scale of the distribution, which controls the spread of the shape

Sampling Method:

Hamiltonian Monte Carlo (HMC), specifically the No-U-Turn Sampler (NUTS) is used as MCMC sampler.

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:

Feature Engineering:

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*. Defined as the variable divided by *Base_Time*.

The original data are highly skewed, and scale (range of the column) varies dramatically between variables. From Figure 1, I projected the data points to *CC1* and *Page_Popularity_Likes*. It is obvious that the data scale varies drastically, and high skewness occurs. So, I decided to perform a log transformation, and normalize the variable to mitigate the serios skewness and largely varying scales.

Figure 1. Data Distribution before 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

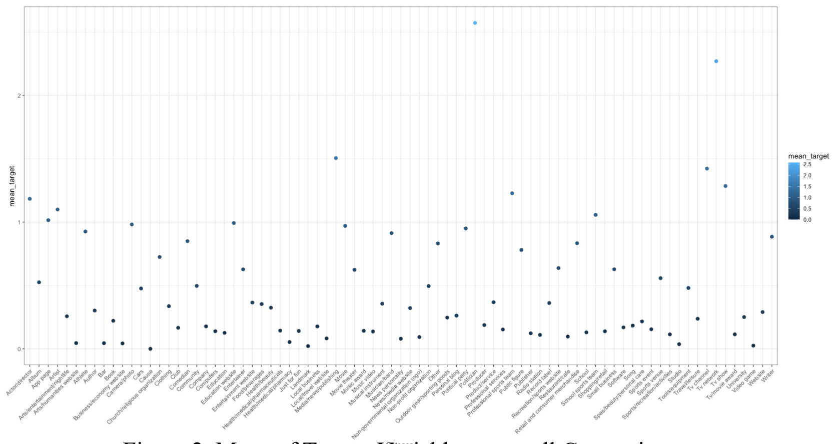


Figure 2. Mean of Target_Variable across all Categories

And by observing different mean *Target_Variable* from the data, shown in Figure 2. 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*. Offset term is also included in the set of the covariates.

Adjustment:

The original training dataset consisted of 199,030 observations, where the negative binomial regression with a weakly informative prior was successfully fitted. When I decided to use a Laplace distribution as the prior, my first step was to apply cross-validation to select the optimal λ parameter, which represents the scale of the Laplace distribution. This process involved initially sampling five λ values from an exponential distribution with a rate of 0.8, and then employing these in a 5-fold cross-validation. However, this entire procedure could not be completed within the 12-hour maximum time limit set by the shared computing cluster (SCC) at Boston University. Consequently, I resorted to sampling a subset of the data using a stratified 10-fold sampling method (19903 observation) to maintain the proportion of the Category. From this subset, I used the first fold as my training data for all the models. Additionally, I set the initial value of the scale parameter, λ, to 1 and enabled the *auto_scale* parameter (setting it to TRUE), allowing the program to automatically adjust the scale for me.

Results

For all the three regressions, priors for intercept follows normal distribution, and prior for covariance applies LKJ distribution on correlation matrix, and gamma(1,1) distribution on squared scale parameter of the decomposed trace of covariance matrix. Table 1 to Table 3 shows the result of estimation.

➢ Poisson regression – weakly informative prior:

Table 1. Poisson Regression - Weakly Informative Prior Estimation

	Posterior Mean	0.05 quantile	0.95 quantile
(Intercept)	-4.0313	-4.3113	-3.7658
CC1_logNorm	-1.1731	-1.2443	-1.1030
CC2_logNorm	2.3020	2.2674	2.3364
CC3_logNorm	0.2122	0.1919	0.2328
CC4_logNorm	-2.8222	-2.8854	-2.7564
CC5	-0.0004	-0.0005	-0.0004
Page_Popularity_Likes_logNorm	0.1182	0.1060	0.1305
Page_Checkins_logNorm	-0.0867	-0.0921	-0.0813
Page_Talking_About_logNorm	0.7561	0.7411	0.7701
Post_Length_logNorm	0.0762	0.0708	0.0817
Post_Share_Count_logNorm	0.4498	0.4435	0.4561
CC2_per_hr_logNorm	-0.6967	-0.7284	-0.6656
CC3_per_hr_logNorm	-0.1772	-0.1905	-0.1642
CC4_per_hr_logNorm	2.2309	2.1974	2.2653

- Negative binomial regression – weakly informative prior:

Table 2. Negative Binomial Regression - Weakly Informative Prior Estimation

	Posterior Mean	0.05 quantile	0.95 quantile
(Intercept)	-3.3089	-3.5653	-3.0496
CC1_logNorm	-1.3006	-1.6989	-0.8895
CC2_logNorm	2.2606	2.1651	2.3565
CC3_logNorm	0.3273	0.2127	0.4413
CC4_logNorm	-3.4951	-3.9295	-3.0846
CC5	-0.0001	-0.0006	0.0005
Page_Popularity_Likes_logNorm	0.2876	0.2249	0.3476
Page_Checkins_logNorm	-0.0158	-0.0515	0.0224
Page_Talking_About_logNorm	0.6594	0.6001	0.7207
Post_Length_logNorm	0.0960	0.0663	0.1263
Post_Share_Count_logNorm	0.7540	0.7186	0.7893
CC2_per_hr_logNorm	-1.5223	-1.6824	-1.3728
CC3_per_hr_logNorm	-0.3612	-0.4567	-0.2639
CC4_per_hr_logNorm	3.5982	3.4305	3.7662

- Negative binomial regression – Laplace distribution prior:

Table 3. Negative Binomial Regression - Laplace Prior Estimation

	Posterior Mean	0.05 quantile	0.95 quantile
(Intercept)	-0.2584	-0.4122	-0.1161
CC1_logNorm	-0.8415	-1.3930	-0.2588
CC2_logNorm	1.5571	1.4588	1.6522
CC3_logNorm	0.2913	0.1871	0.3953
CC4_logNorm	-1.4331	-2.0029	-0.8751
CC5	-0.0002	-0.0006	0.0001
Page_Popularity_Likes_logNorm	0.2476	0.1886	0.3075
Page_Checkins_logNorm	-0.0572	-0.0907	-0.0233
Page_Talking_About_logNorm	0.4257	0.3653	0.4860
Post_Length_logNorm	0.0714	0.0426	0.0992
Post_Share_Count_logNorm	0.5729	0.5400	0.6072

CC2_per_hr_logNorm	-0.5644	-0.7066	-0.4268
CC3_per_hr_logNorm	-0.1762	-0.2632	-0.0916
CC4_per_hr_logNorm	1.6605	1.5131	1.8195

Likes, revisiting, shares, and the average numbers of hourly comment posted withing the first 24 hours after the post was published contribute the most to the comment volume and the credibility is supported by the 95% posterior interval.

Appendices 2 to 5 include the Posterior Predictive Check and MCMC diagnostics, demonstrating a strong fit for most of the data. The Prior versus Posterior plot clearly illustrates the higher peak of the Laplace distribution in comparison to a normal distribution. Additionally, the autocorrelation plots indicate that all samples of fixed effects converge to approximately zero.

Prediction:

Evaluation Metrics:

Table 4 shows the predictive result of the four performance metrics for each model.

Table 4. Prediction Performance Metrics

Model	MSE	MAPE	RMSE	MAE
Poisson regression – weakly informative prior	11525.61	0.2266801	107.3574	25.92
Negative Binomial – weakly informative prior	11292.73	1.215907	106.2673	25.681
Negative Binomial – Laplace distribution	11626.28	0.3794737	107.8252	26.398

The use of a Laplace prior with a negative binomial sampling model has significantly reduced the Mean Absolute Percentage Error (MAPE), from 1.22 to 0.38. Interestingly, the choice between a Poisson and a Negative Binomial sampling model shows only a minor difference in predictive performance. However, this apparent similarity may be misleading. The high volatility in the results, primarily due to long-tailed predictions, can lead to substantial differences in the performance metrics of the three regression models, despite all of them fitting well.

Conclusion:

In conclusion, this research presents a detailed analysis of three regression models used to predict comment volumes on Facebook. It highlights the superior performance of the Laplace distribution in terms of Mean Absolute Percentage Error (MAPE) when compared to models using weakly informative priors. Additionally, the study examines the impact of different sampling models, specifically Poisson and Negative Binomial, revealing a minor difference in predictive performance despite the presence of overdispersion in the data. The comprehensive fit of the models, as detailed in Appendix 2, suggests that the variance in predictive performance is likely attributable to the significant skewness of the data, resulting in considerable variability in predictions.

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.

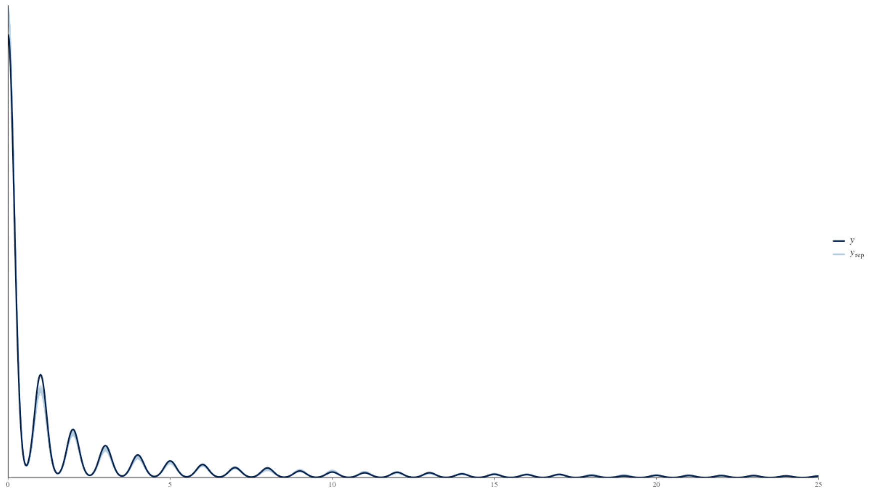
Appendix 1. Training Datasets Summary Statistics

Statistic	N	Mean	St.Dev.	Min	Max
Page_Category	19,903	24.337	20.096	1	106
Page_Popularity_Likes	19,903	1277042	6113990	36	486972297
Page_Checkins	19,903	4730.547	20805.73	0	186370
Page_Talking_About	19,903	44415.59	110264.3	0	6089942
CC1_Min	19,903	0.39	8.134	0	486
CC1_Max	19,903	486.865	540.365	0	2442
CC1_Avg	19,903	56.137	88.261	0	1256.517
CC1_Median	19,903	35.596	70.778	0	1404
CC1_Std	19,903	68.188	83.153	0	762.358
CC2_Min	19,903	0.065	1.606	0	113
CC2_Max	19,903	382.55	441.442	0	2119
CC2_Avg	19,903	21.85	36.006	0	577.744
CC2_Median	19,903	7.222	19.526	0	565
CC2_Std	19,903	40.492	51.498	0	457.966
CC3_Min	19,903	0	0	0	0
CC3_Max	19,903	380.685	430.077	0	2095
CC3_Avg	19,903	20.14	32.809	0	505.828
CC3_Median	19,903	4.899	13.352	0	405
CC3_Std	19,903	40.837	53.67	0	613.8
CC4_Min	19,903	0.39	8.129	0	486
CC4_Max	19,903	435.989	492.768	0	2184
CC4_Avg	19,903	52.95	82.451	0	1084.242
CC4_Median	19,903	33.9	66.021	0	1105
CC4_Std	19,903	63.511	77.382	0	680.962
CC5_Min	19,903	-326.194	380.302	-2038	0
CC5_Max	19,903	378.347	438.562	0	2119
CC5_Avg	19,903	1.71	9.604	-150.333	272.385
CC5_Median	19,903	-2.134	11.219	-165	521
CC5_Std	19,903	56.625	75.328	0	771.339
CC1	19,903	56.775	143.459	0	2410
CC2	19,903	21.444	73.931	0	1852
CC3	19,903	20.811	78.856	0	1738
CC4	19,903	53.433	133.429	0	2082
CC5	19,903	0.634	95.073	-1450	1852
Base_Time	19,903	35.661	21.013	0	72

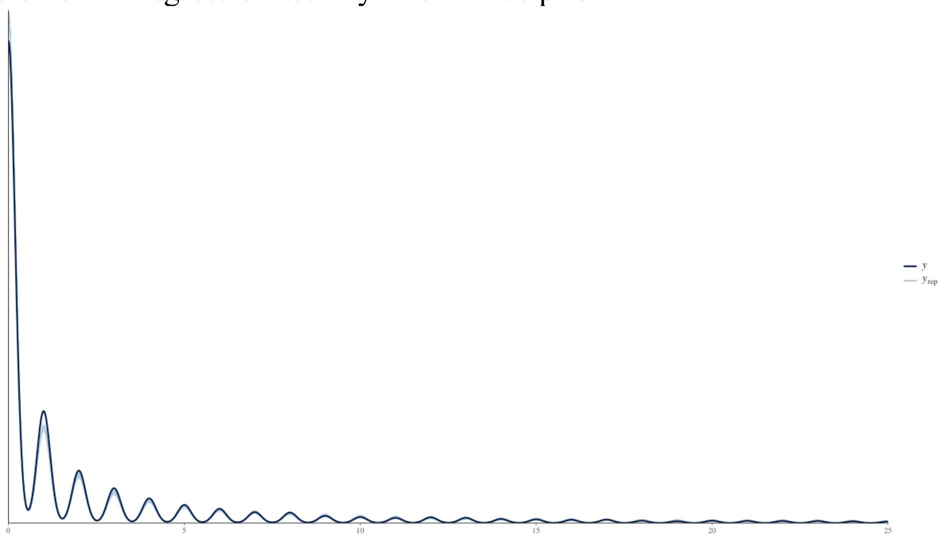
Post_Length	19,903	162.699	379.425	0	14497
Post_Share_Count	19,903	122.759	1148.049	1	144860
Post_Promotion_Status	19,903	0	0	0	0
H_Local	19,903	23.784	1.856	1	24
Post_Published_Weekday_40	19,903	0.123	0.329	0	1
Post_Published_Weekday_41	19,903	0.143	0.35	0	1
Post_Published_Weekday_42	19,903	0.151	0.358	0	1
Post_Published_Weekday_43	19,903	0.157	0.363	0	1
Post_Published_Weekday_44	19,903	0.145	0.352	0	1
Post_Published_Weekday_45	19,903	0.146	0.353	0	1
Post_Published_Weekday_46	19,903	0.136	0.343	0	1
Base_DateTime_Weekday_47	19,903	0.14	0.347	0	1
Base_DateTime_Weekday_48	19,903	0.132	0.338	0	1
Base_DateTime_Weekday_49	19,903	0.138	0.344	0	1
Base_DateTime_Weekday_50	19,903	0.147	0.354	0	1
Base_DateTime_Weekday_51	19,903	0.159	0.366	0	1
Base_DateTime_Weekday_52	19,903	0.145	0.352	0	1
Base_DateTime_Weekday_53	19,903	0.14	0.347	0	1
Target_Variable	19,903	7.044	33.312	0	1429
CC2_per_hr	19,903	2.042	10.666	0	757
CC3_per_hr	19,903	0.523	2.047	0	49.543
CC4_per_hr	19,903	2.735	10.893	0	757
CC1_logNorm	19,903	-0.0002	0.999	-1.413	2.943
CC2_logNorm	19,903	-0.007	0.994	-0.929	3.782
CC3_logNorm	19,903	0.005	1.006	-0.767	3.828
CC4_logNorm	19,903	-0.0004	0.999	-1.397	2.903
Page_Popularity_Likes_logNorm	19,903	-0.007	1.001	-3.699	3.352
Page_Checkins_logNorm	19,903	0.002	1.001	-0.643	2.838
Page_Talking_About_logNorm	19,903	0.001	1	-2.717	2.364
Post_Length_logNorm	19,903	0.0003	0.991	-2.345	3.099
Post_Share_Count_logNorm	19,903	0.001	1.004	-1.112	4.901
CC2_per_hr_logNorm	19,903	-0.009	0.988	-0.542	7.642
CC3_per_hr_logNorm	19,903	0.007	1.016	-0.465	7.98
CC4_per_hr_logNorm	19,903	-0.004	0.995	-0.778	6.862

Appendix 2. Posterior Predictive Density Plot

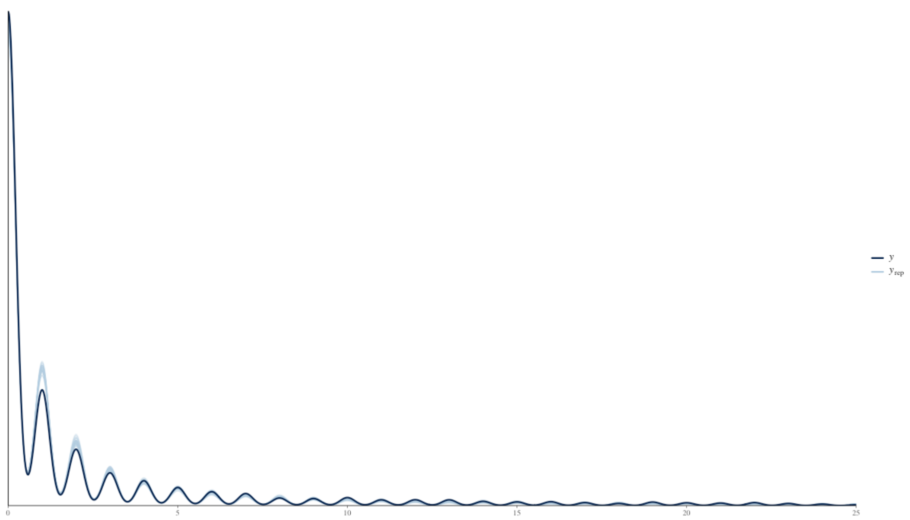
- Poisson regression -weakly informative prior (Only showing the xaxis between 0 to 25)



- Negative binomial regression-weakly informative prior

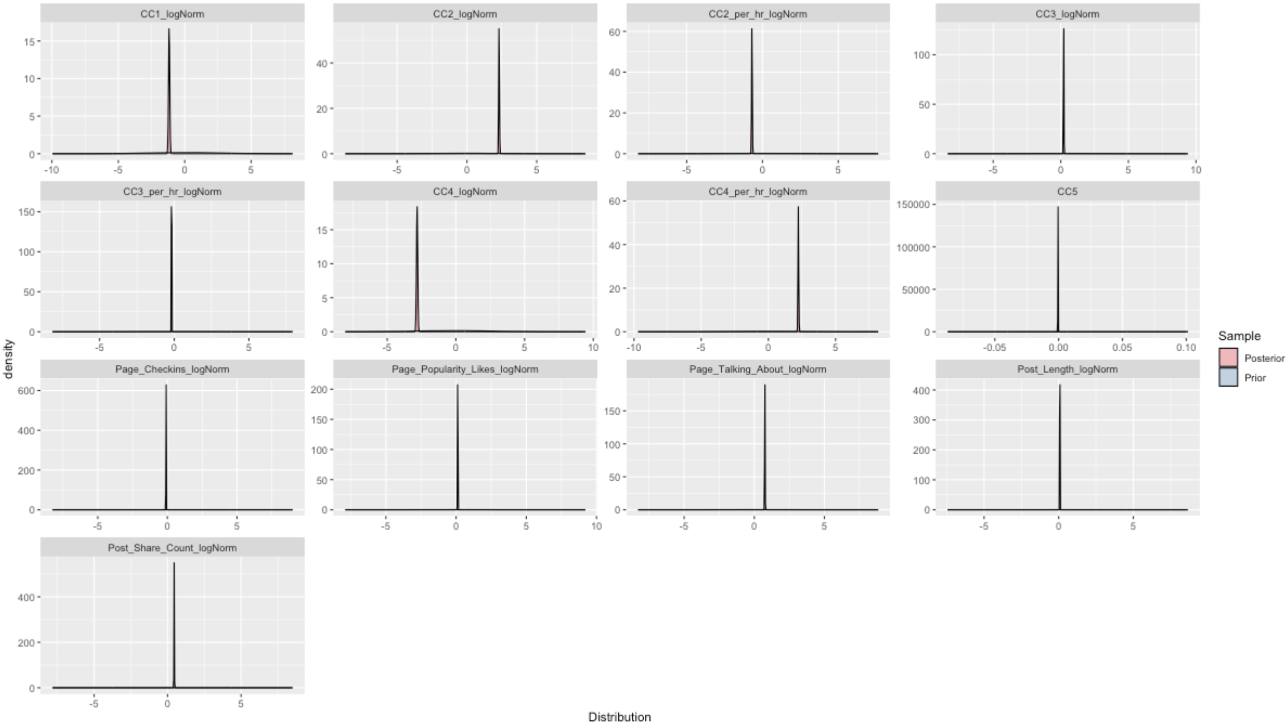


- Negative binomial regression- Laplace prior

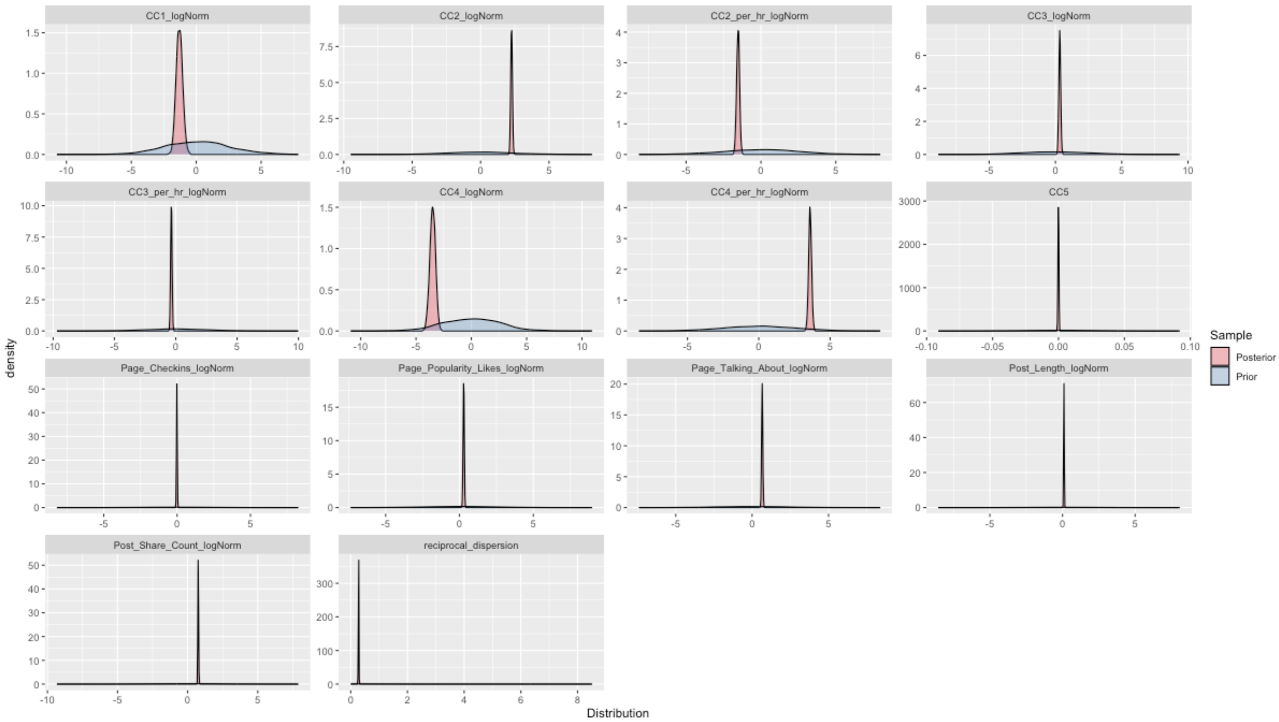


Appendix 3. Prior versus Posterior Sample Comparison

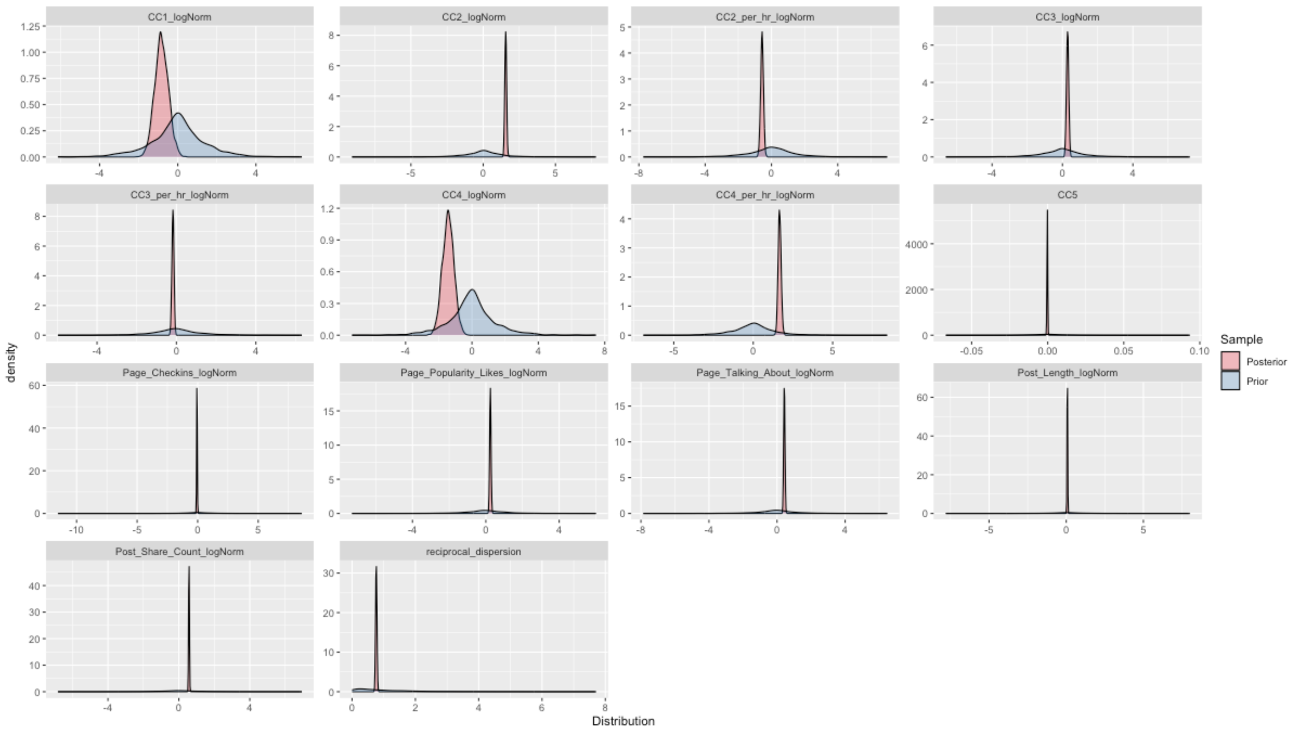
- Poisson regression -weakly informative prior (Only showing the xaxis between 0 to 25)



- Negative binomial regression-weakly informative prior

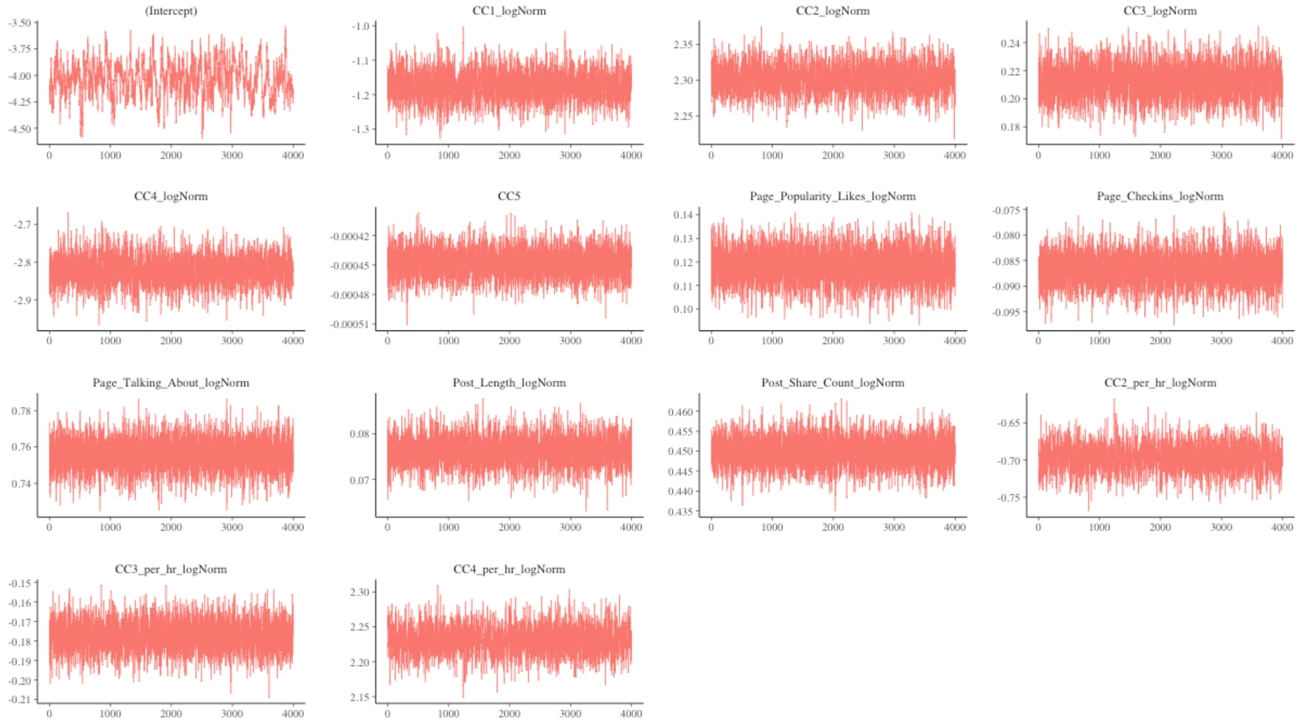


- Negative binomial regression- Laplace prior

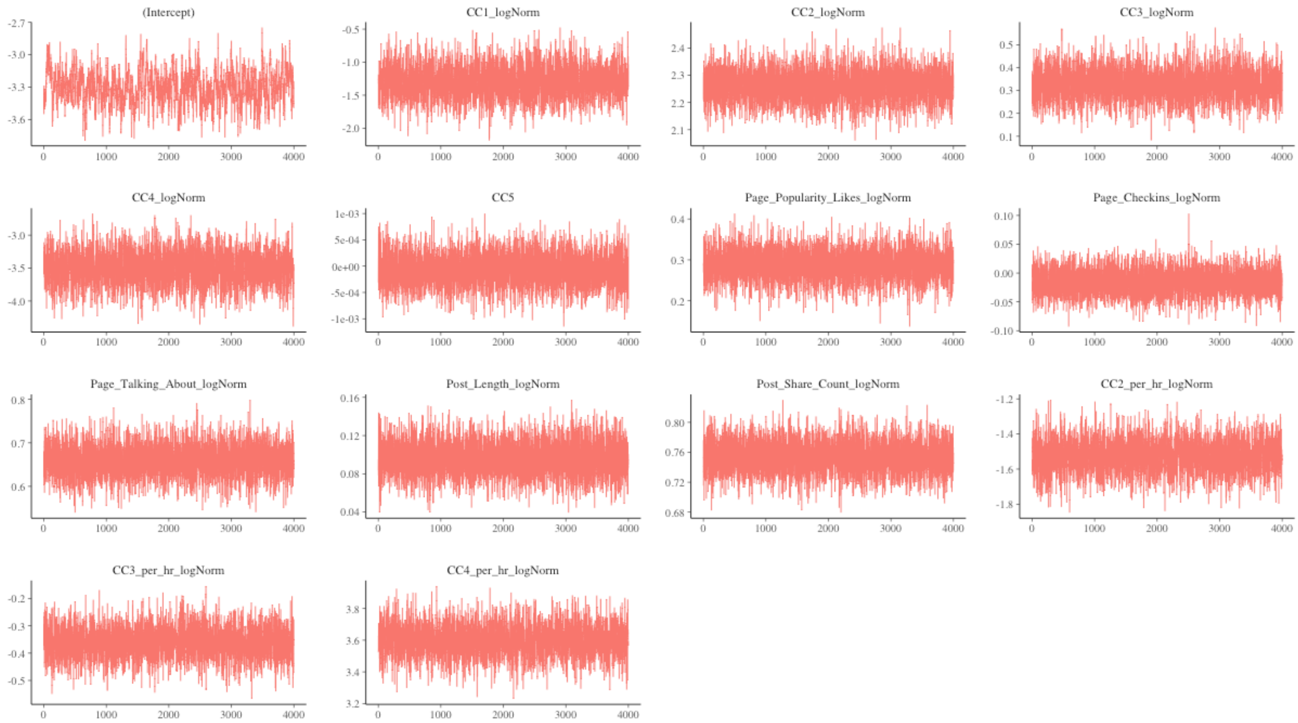


Appendix 4. MCMC Diagnostics – Trace Plot

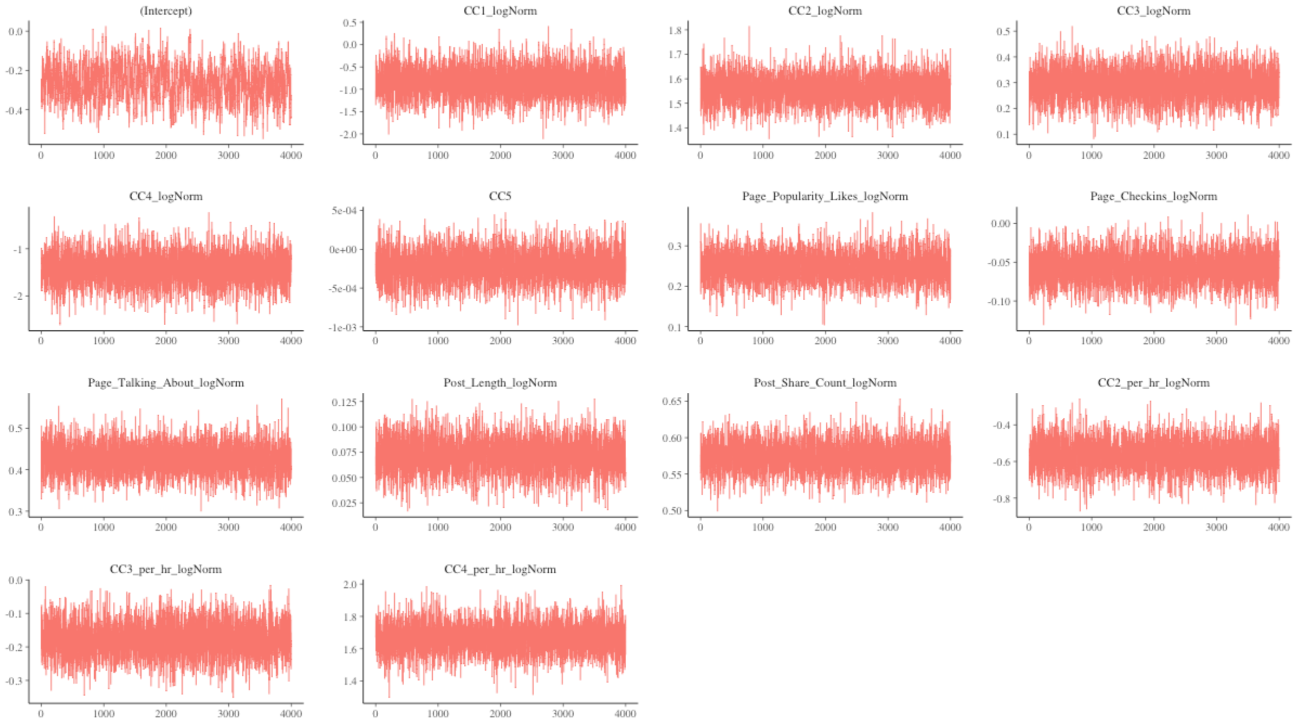
➢ Poisson regression -weakly informative prior



➢ Negative binomial regression-weakly informative prior

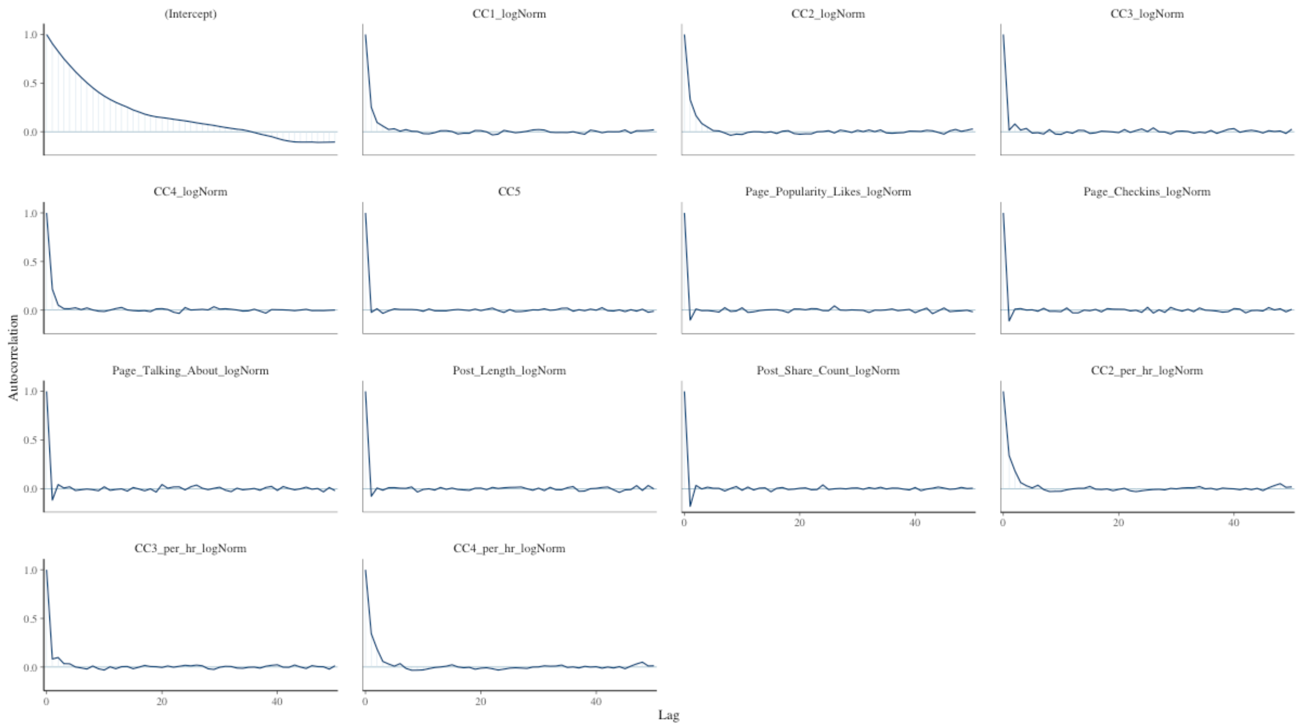


- Negative binomial regression- Laplace prior

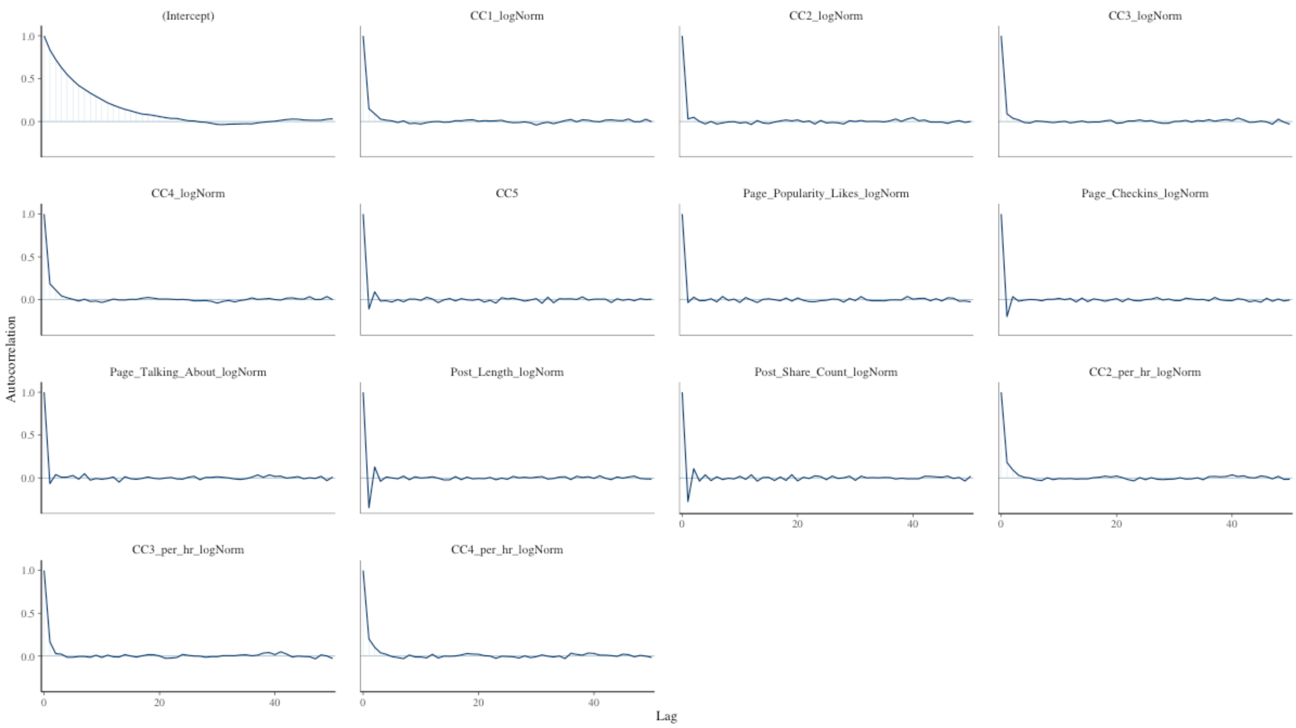


Appendix 5. MCMC Diagnostics – Acf Plot

- Poisson regression -weakly informative prior



- Negative binomial regression-weakly informative prior



- ➢ Negative binomial regression- Laplace prior