Generalize Linear Models for Flight Prediction

Steven Lan
slan4@jh.edu

Xiyu Li
xiyuli027@gmail.com

Xiaoya Huang
xhuang98@jh.edu

Abstract

The effective prediction of flight delays is a widely discussed topic in both aviation industry research and practical applications for individuals. The accuracy of these predictions heavily relies on the selection of variables and the availability of data. While many studies have utilized regression models and other predictive techniques, our research introduces a novel approach that integrates Bayesian methods to leverage prior knowledge of influential variables, thereby enhancing the accuracy of flight delay predictions. Specifically, we use logistic regression and mixed-effects models not only to assess the likelihood of flight delays but also to estimate the expected duration of these delays. This dual-focus approach enables us to effectively apply our domain knowledge within a Bayesian framework, thereby improving the specificity and accuracy of our predictions. Additionally, by employing Bayesian inference, we aim to provide insights into which variables are most significant according to our model.

1 Introduction

Predicting flight delays requires sophisticated models capable of handling the complex and variable factors that affect flight schedules. Our project utilizes data from the Bureau of Transportation Statistics, a comprehensive U.S. government source that records detailed flight operations. Such data forms the empirical backbone of our models.

Furthermore, the aviation industry's strict regulations, like the maximum allowable wind speed for departures, significantly influence scheduling. These rules help establish informed priors in our Bayesian framework, enhancing our model's predictive accuracy by integrating these operational thresholds and amendments to airline policies.

2 Related Work

Predicting flight delays is a widely researched topic, with many studies traditionally employing various regression models due to their efficacy in handling binary outcomes across numerous variables. While regression models are foundational, recent research has expanded to include a range of classification methods. A notable study compared the performance of logistic regression, XGBoost, Random Forest, and Multilayer Perceptron (MLP) on a specific dataset. In this comparison, XGBoost emerged as the superior model, demonstrating the highest accuracy.

However, when applying XGBoost to our dataset, it achieved only a modest accuracy of 0.55, which did not meet our expectations. This has directed our research towards leveraging a Bayesian approach. Bayesian models are particularly promising due to their ability to incorporate prior knowledge into the estimation process, potentially enhancing the accuracy of predictions. We hypothesize that the integration of prior information in our Bayesian model will yield more accurate and robust predictions of flight delays compared to the other methods.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

3 Methodology

Our research aims to address two principal questions within the domain of flight delay prediction using Bayesian statistical methods.

Influence of Variables on Departure Delays The first segment of our study seeks to understand the relationship between selected variables and the likelihood of a flight's departure being delayed by more than 15 minutes. Initially, we identify critical variables based on airline guidelines and preliminary exploratory data analysis (EDA). This step ensures that the factors included in our model are relevant and have a potential impact on departure delays.

Following the selection of pertinent variables, we will employ logistic regression to model the probability of delays. This model will help us discern the most and least significant predictors of delay. To refine our model and further validate the importance of each variable, we will conduct model selection procedures.

Subsequently, we will implement Markov Chain Monte Carlo (MCMC) simulations to derive posterior distributions for the parameters of interest. The goal is to achieve a substantial effective sample size that assures the stability and reliability of our Bayesian inference.

Impact of Airline Operations on Delay Duration The second part of our research focuses on whether the specific airline operating the flight significantly affects the actual delay duration. To explore this, we will use a mixed-effects model that assumes departure delay times follow a zero-inflated Poisson distribution. This approach allows us to account for the excess zeros (flights not delayed) and the over-dispersion in delay times.

In this model, the random effects will be modeled to capture the variability across different airlines, with a particular focus on the intercepts of these airlines. Through MCMC, we aim to estimate these effects accurately and assess the differences among airlines regarding their impact on delay times.

Progress Plan Our initial focus will be on addressing the first research question. This priority is set to establish a strong foundation in understanding how individual variables influence the likelihood of delays. Time permitting, we will extend our analysis to the hierarchical linear model to explore the second research question comprehensively.

3.1 Dataset

Our dataset combines 2 datasets, flight data which we derive from the Bureau of Transportation Statistics, and weather data which we derive from Visual Crossing. The flight data includes 93 features, ranging from June 2018 to May 2022 of routes LAX-LAS and LAX-JFK. The weather data merges to be the corresponding daily data of each flight, with 22 features.

The last 23 columns in original flight dataset has no entries. After removing the empty columns, we started from finding the missing values. As shown in Figure 1, there are numerous missing values in 9 columns. For the first 4 features, the number of missing values equals the number of data in the dataset. It indicates there are no entries in these features, and the features are unhelpful for delay prediction. We removed the 4 columns.

Missing Values in the dataset

- **CancellationCode**: Missing values in CancellationCode are reasonable. By finding the number of canceled flights, we found every canceled flight has its cancellation code. Otherwise, there are no entries.
- **Five delay reason features**: All missing values in these 5 features are in the same rows.
- **Delayed flight without delay reason:** By checking the count of canceled/not delayed flights, we found there are around 17000 delayed flights missing delay reasons. But all flights were delayed over 15 minutes and only 1862 within 15 minutes(total 21074 flights) have delay reasons stated. In our analysis, we consider only flights delayed over 15 minutes as delayed flights. The missing values are reasonable under the situation of small delay minutes.

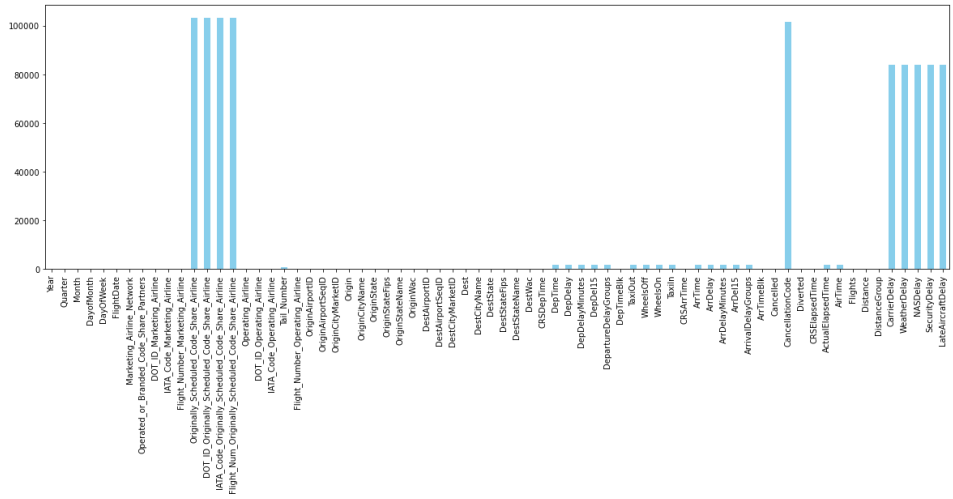

Figure 1: count of NA for each feature in flight dataset

Core feature with Missing Value: DepTime, DepDelay, DepDelayMinutes, DepDel15, ArrTime, ArrDelay

All departure features have missing values at the same rows. All arrival features have missing values at the same rows, except there are 46 rows that have only ArrTime entries. In order to find the reason for these missing values, the plot "Count of NA v.s. Date" is graphed.

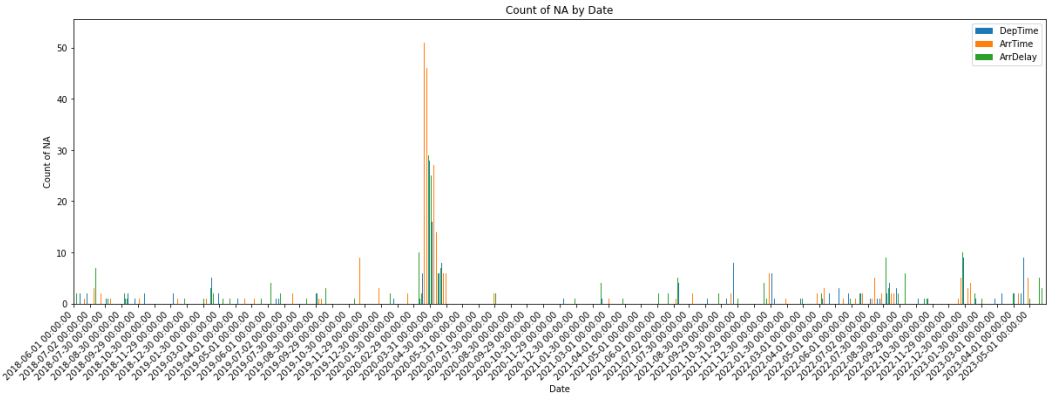

Figure 2: NA Count of Core Feature by Date

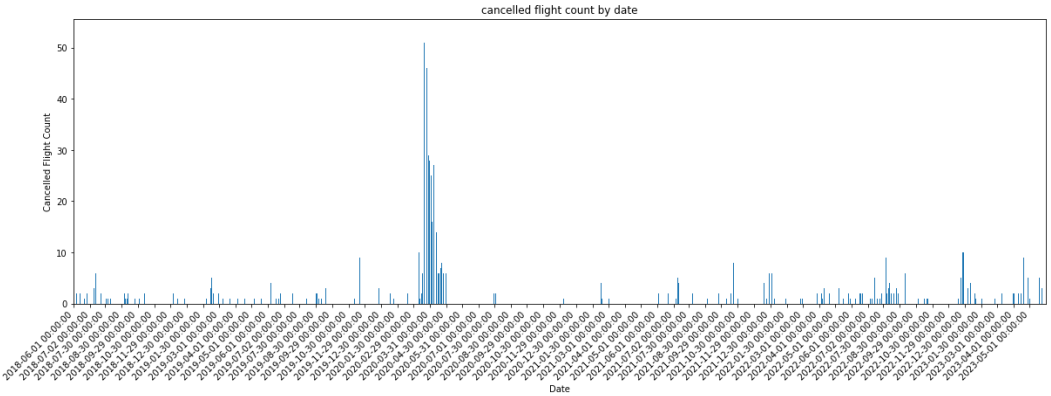

Figure 3: Cancelled Flight Count by Date

Comparing Figure2 and Figure3, cancellation due to COVID19(mainly during Year 2020) is the main reason for missing values in arrival and departure features. Since cancelled flights gives no information for the prediction, all cancelled flights are dropped. And due to the high rate of cancellation, the total number of flights and delay rate in year of 2020 reduced significantly. While the delay rate of other years are in the range of 0.171 to 0.239, year 2020 has 0.079. As COVID19 is not a common situation, flights in Year 2020 are dropped.

Based on the definition of delay reason given by the original data website, only flights with delay reason of type NAS and Weather can be explained with the features we have in our dataset(NAS delay include: weather conditions, airport operations, heavy traffic volume, air traffic control). So we further cleaned the data by selecting the flights with delay reason of only NASDelay and WeatherDelay. The result dataset has only 252 rows left. We randomly selected the same number of undelayed flights and merged them, resulting in a 502 rows dataset that we used in later modeling.

Exploration Data Analysis: We conducted a mapping of flights departing from different airports to assess the potential impact of origin airports on delays. The histogram illustrates that the percentage of delays at JFK is notably higher compared to those at LAS and LAX.

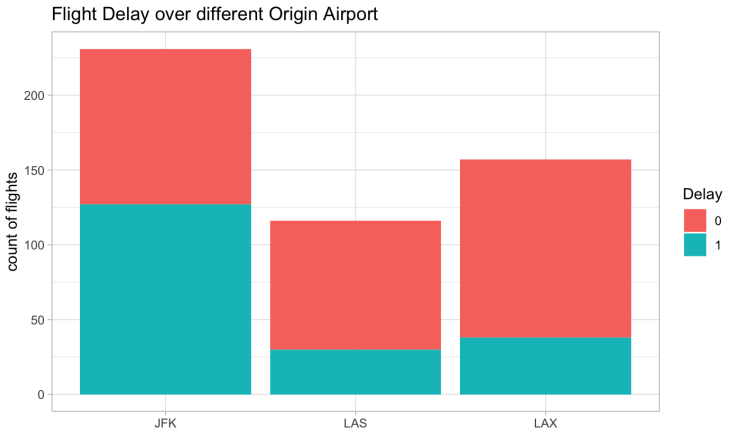

Figure 4: Histogram of Flights over Departure Airport(Origin)

In order to visualize the influence of weather conditions on flight delays, we constructed a histogram. Although our dataset contains few instances labeled as cloudy days, it is evident that days with rain or snow may exhibit a strong positive relationship with flight delays.

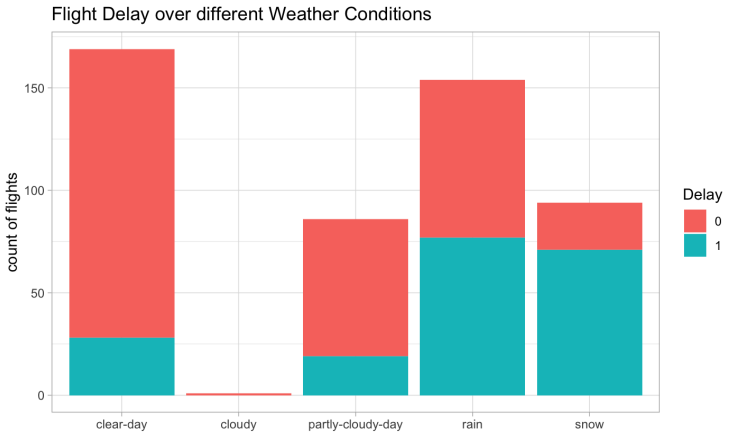

Figure 5: Histogram of Flights over Weather Condition(icon)

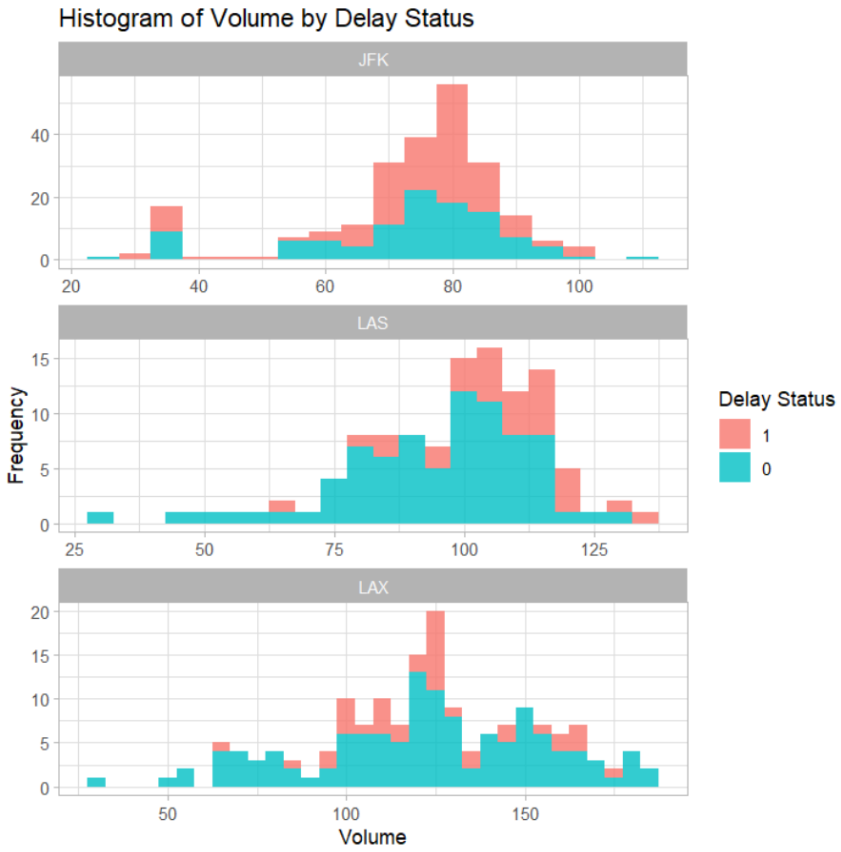

Figure 6: Enter Caption

3.2 Model

3.2.1 Delay Analysis

Frequentist Approach As the result of EDA, volume, Origin, snowdepth and icon are considered as the predictors. The result is shown in Figure 7. The model has accuracy of 0.723 and correlation of prediction and true y of 0.436.

	Estimate <dbl>	Std. Error <dbl>	z value <dbl>	Pr(>|z|) <dbl>
(Intercept)	-2.296076472	0.560645641	-4.0954148	4.214131e-05
volume	0.008746685	0.005900864	1.4822718	1.382680e-01
OriginLAX	-0.561023151	0.485492676	-1.1555749	2.478551e-01
OriginLAS	-0.060342184	0.397564129	-0.1517797	8.793607e-01
iconsnow	2.710817835	0.424451971	6.3866303	1.695809e-10
iconrain	1.701678498	0.323757205	5.2560328	1.471960e-07

Figure 7: Result of Frequentist Approach

Bayesian Approach using stan package Prior choice: All beta is generated from Normal distribution. The intercept has mean equal to log odd of delay probability, and standard deviation equal to absolute half mean. The remaining beta use flat prior with mean = 0 and sd = 2.5. The result is shown in Figure 8. The model has accuracy of 0.6870748 and correlation of prediction and true y of 0.2972822, slightly lower than the Frequentist Approach.

```
Estimates:
              mean   sd   10%   50%   90%
(Intercept) -1.9    0.6 -2.7  -1.9  -1.2
snowdepth    0.0    0.1 -0.1   0.0   0.1
volume       0.0    0.0  0.0   0.0   0.0
OriginLAX   -0.7    0.5 -1.3  -0.7  -0.1
OriginLAS   -0.2    0.4 -0.7  -0.2   0.4
iconrain     1.1    0.3  0.6   1.0   1.5
iconsnow     2.3    0.5  1.7   2.3   2.9
```

Figure 8: Result for using stan package

Bayesian Approach with handcode:

Methodology We split the dataset into training and test sets with a ratio of 0.7 for training and 0.3 for testing. The MCMC simulation was conducted on the test dataset, generating 10,000 samples. A flat prior was utilized for the regression coefficients, and the initial values for the coefficients were set to those obtained from the frequentist model. We employed a multivariate normal distribution as the proposal distribution for generating new coefficients, with a scaling parameter $c = 5$.

Effective Sample Size Consideration Despite the promising results, it's important to note that the effective sample size for each beta was around 400. This lower-than-desired effective sample size was primarily due to the high autocorrelation observed in the beta values. This autocorrelation stemmed from the use of the Metropolis-Hastings acceptance criterion during the MCMC simulation.

Column <chr>	Mean <dbl>	SD <dbl>	X10. <dbl>	X50. <dbl>	X90. <dbl>
snowdepth	0.01853	0.07965	-7.91e-02	0.01419	0.1189
volume	0.00796	0.00631	-1.43e-05	0.00769	0.0161
OriginLAX	-0.40614	0.51101	-1.06e+00	-0.39564	0.2519
OriginLAS	0.76650	0.46386	1.68e-01	0.76265	1.3766
iconrain	1.97380	0.35980	1.52e+00	1.96425	2.4368
iconsnow	2.62673	0.49375	2.01e+00	2.59766	3.2786
(Intercept)	-2.44809	0.62736	-3.26e+00	-2.41647	-1.6871

Figure 9: Result from Hand code MCMC

Results The model achieved an accuracy of 0.788 on the test data, indicating its predictive capability. Additionally, the correlation between the model's predictions and the true values (denoted as y) was found to be 0.558, which represents a notable improvement compared to previous models.

The coefficients obtained from our hand-coded model provide valuable insights into the relationship between chosen predictors and flight delay (DepDel15). Our analysis reveals that:

- Weather Conditions: Both rainfall and snowfall show a positive relationship with flight delay, indicating that adverse weather conditions contribute to increased delays.
- Origin Airport Significance: The origin airport variable plays a significant role in predicting flight delays, suggesting that factors specific to each airport influence the likelihood of delays.
- Snow Depth and Volume: Notably, snow depth and volume exhibit positive beta means, further highlighting their impact on flight delays during snowy conditions.

These findings align with our expectations and suggest that our model captures meaningful relationships between predictors and flight delay outcomes.

3.2.2 Delay Duration Analysis

Data For this analysis, we utilize a small subset of the dataset, comprising 800 data points in total. This subset encompasses an equal distribution among various operating airlines, with each represented by 100 individual data entries. Figure illustrates the distribution of departure delay minutes across these airlines.

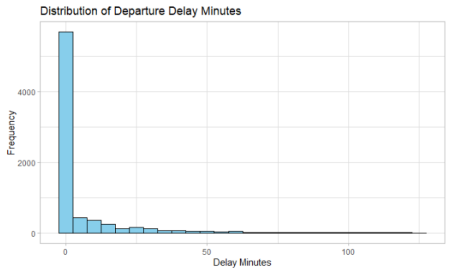

Figure 10: Delay Distribution

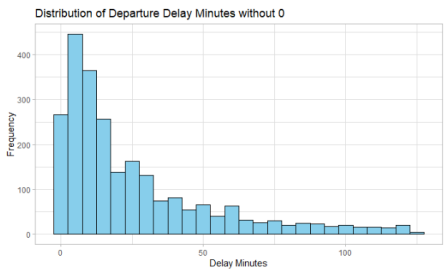

Figure 11: Delay Distribution without 0

Figure 12: distribution of departure delays

Frequentist Approach Instead of predicting the occurrence of flight delays, our project is also interested in identifying the variables that most significantly influence the minutes of departure delay. To achieve this, we employed a mixed-effects model with the operating airline serving as the group-level predictor. Our initial step involved an examination of the raw data on departure delay minutes, which suggested a distribution akin to Poisson. Consequently, we implemented several non-Bayesian hierarchical linear models (HLM) initially, including an intercept-only model, a categorical data-only model, and a full model. Model selection was guided by examining the t-values from these preliminary models, ultimately leading to the selection of our final model. Comparative plots were generated to illustrate the model fits. Notably, these plots revealed inadequate fitting, prompting us to address the excessive number of zeros in the departure delay data. After excluding these zero-delay instances, we repeated the analysis. However, as depicted in the subsequent graphs, the refined model still struggled to accurately predict the minutes of departure delay, indicating the need for further model adjustments or alternative analytical approaches.

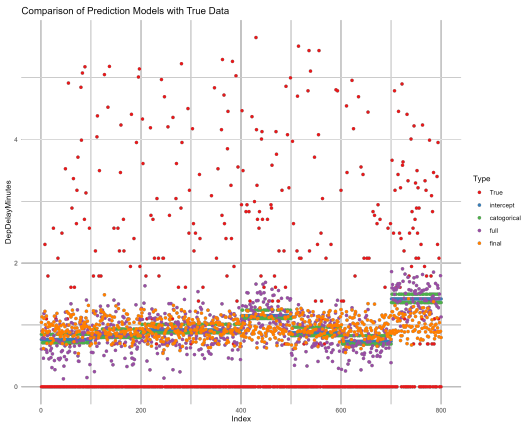

Figure 13: Frequentist Approach with full data

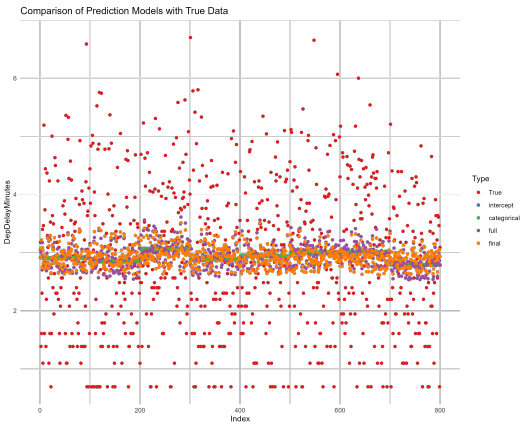

Figure 14: Frequentist Approach without 0s

While the frequentist approach did not adequately address model fitting, it yielded some notable findings. Specifically, the analysis indicated that the operating airline had minimal influence on departure delay minutes. In contrast, the month of operation demonstrated a more significant impact on the duration of departure delays. Given the limitations of the frequentist models, we opted to explore Bayesian methods.

Bayesian Approach Consistent with our earlier assessment that departure delay minutes follow a Poisson distribution, we opted for weak priors to minimize the influence of prior assumptions on the distribution of the variables. Accordingly, we selected a normal prior with a mean of zero and a standard deviation of 25 for the continuous variables. For categorical data, we utilized T-distributions. Following the establishment of these priors, we conducted MCMC simulations to obtain the posterior distributions. The results of these simulations, which illustrate the influence of various factors on departure delay minutes, are depicted in the accompanying plots. We run the ACF and see some of the variables have high correlations, and the effective sample size is also low.

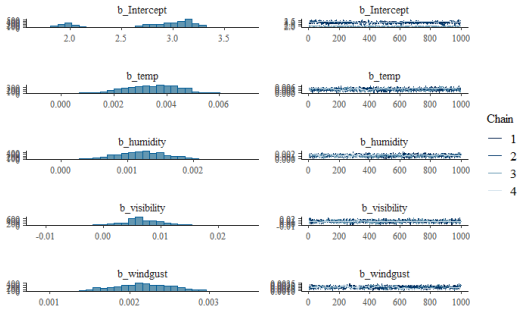

Figure 15: trace plot for continuous variables

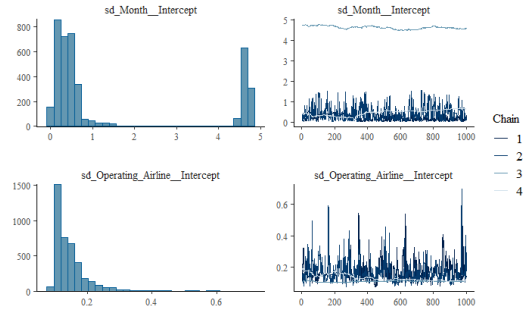

Figure 16: trace plot for catogorical variables

```
Multilevel Hyperparameters:
~Month (Number of levels: 2)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.74      0.57     0.10     1.65 1.72        6       13

~Operating_Airline (Number of levels: 8)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.32      0.14     0.16     0.66 1.57        7       41

Regression Coefficients:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept      2.50      0.65     1.37     3.37 1.59        7       33
temp           0.02      0.00     0.02     0.02 1.01      245      409
humidity       0.01      0.00     0.00     0.01 1.25       12       42
visibility     0.00      0.00    -0.00     0.01 1.17       17       72
windgust       0.00      0.00     0.00     0.00 1.01      344      541
```

Figure 17: Result for HLM

Results The results from our Bayesian analysis are quite revealing. The confidence intervals for the intercepts and the variable 'Month' demonstrate the most substantial influence on departure delay minutes. This suggests that the variables included in our dataset are largely insufficient to account for variations in departure delay minutes, with the operating airline showing negligible or no effect. Therefore, we conclude that the timing of the flight, specifically whether it occurs in peak or low months, plays a critical role in predicting departure delays. This insight highlights the importance of temporal factors over operational ones in managing and anticipating flight delays.

4 Reproducible

This work can be reproducible in this Github Repo.

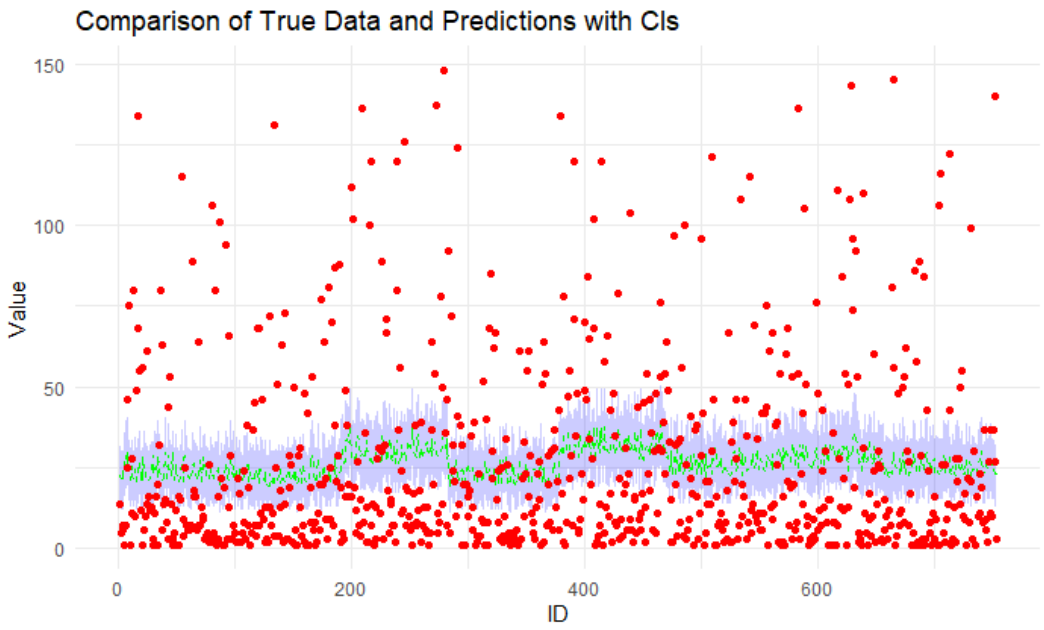

Figure 18: Compare true and predict value

References

[1] Lui, Go Nam and Nguyen, Chris HC and Hui, Ka Yiu and Hon, Kai Kwong and Liem, Rhea Patricia, Investigation on Weather- and Trajectory-Based Features to Improve Airspace-Specific Aircraft Arrival Transit Time Prediction.