STATS 551 Project Report: Flight Fare Prediction through Bayesian Linear Regression and Bayesian Structural Time Series Analysis

Group 21: Zihan Cui, Xingjian Zhang, Yushan Yang

December 2022

Github URL: https://github.com/yys1234/STATS551-Project-Group21

1 Introduction

Travelling to Miami is a popular holiday plan for UMich students and the most common transportation choice for them is by air. How to purchase a cost-effective flight ticket is a great concern for students. In this project, we are interested in predicting flight fare from Detroit Metro Airport (DTW) to Miami International Airport (MIA). Besides, we want to find out the relation between flight fare and other predictors through Bayesian linear regression and use Bayesian structural time series to analyze the time effect in the data. As a result, Bayesian linear regression with g-prior gave us a $48.1 root mean square error (RMSE) with each parameter significantly affected the flight fare and model selection reduced the RMSE to $45.3. Bayesian structural time series captured the trend and seasonality from the data and successfully forecast an increase pattern in the averaged flight fare each day in last 30 days with a $25.34 RMSE.

2 Methods

The problem for this project is to predict flight fare but the difficulty lies in the fact that flight fare observation is not independent. Therefore, we tried two ways to simplify the problem. First we transformed date variables into some categorical variables and used Bayesian linear regression to do prediction. We also did average of flight fare each day and applied Bayesian structural time series analysis on this univariate variable to figure out the time effects.

2.1 Bayesian Linear Regression (LR) Model

Bayesian linear regression is widely used to determine the posterior of model parameters and generate the probability distribution of the response variable. Suppose our linear model is $y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$, where $\boldsymbol{y}$ is the response, X is the matrix of predictors, $\boldsymbol{\beta}$ is a vector of unknown coefficients and ϵ is the residuals ($\epsilon \sim N(0, \sigma^2)$). The posterior distribution formula is:

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}, \boldsymbol{y}) = \frac{p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)p(\boldsymbol{X}, \boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{X}, \boldsymbol{y})}.$$

which can be calculated given certain priors.

2.1.1 g-prior

The g-prior Bayesian linear regression is motivated by making the distributions of interest remain invariant to changes in parameterization of the model [1]. One possible solution is to set the prior for $\boldsymbol{\beta} \sim N_p(\mathbf{0}, g\sigma(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ and for $\sigma^2 \sim inverse - gamma(\nu_0/2, \nu_0\sigma_0^2/2)$. The corresponding posterior will be:

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{X}, \boldsymbol{y} \sim N_p(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{OLS}, \frac{g}{g+1}\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

$$\sigma^2|\boldsymbol{X}, \boldsymbol{y} \sim inverse - gamma((\nu_0+n)/2, (\nu_0\sigma_0^2 + SSR_g)/2)$$

where $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is the ordinary least squares estimate of $\boldsymbol{\beta}$, $SSR_g = \boldsymbol{y}^T(\boldsymbol{I} - \frac{g}{g+1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y}$ and n is the sample size. The posteriors could be obtained by Monte Carlo samples.

2.1.2 Bayesian linear regression model selection

Sometimes the response variable and predictors do not have a linear relationship and a common way to deal with it is to take the interaction and square terms into account. This will result in a large number of regressors in the model which may be irrelevant to the response variable and cause overfitting. To deal with this problem, model selection is necessary. The Bayesian model selection method selects model by adding a new vector of unknown parameters $\boldsymbol{z}$:

$$y = z_1\beta_1 x_1 + ... + z_p\beta_p x_p + \epsilon$$

where z_j could be 0 or 1 and p is the total number of regressors. The posterior for z_j is:

$$Pr(z_j = 1|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{z}_{-j}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}_{-j}, z_j = 1)}{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}_{-j}, z_j = 0)}$$

We could then use Gibbs sampling procedure to obtain all parameters.

2.2 Bayesian Structural Time Series (BSTS) Model

A general *structural time series model* has three components: a trend component that captures the general trend of the time series, a seasonal component that captures the repeated patterns within the time series, and a regression component that captures the impact of different covariates [2].

Suppose we have a dataset $D = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_T}, y_T)\}$, where $T = |D|$ is the number of timestamps present in the dataset. A structural time series model assumes the following data generation schemes:

$$\begin{aligned} y_t &= \underbrace{\mu_t}_{\text{trend}} + \underbrace{\gamma_t}_{\text{seasonal}} + \underbrace{\boldsymbol{\beta}^T\boldsymbol{x_t}}_{\text{regression}} + \epsilon_t \\ \mu_t &= \mu_{t-1} + \delta_{t-1} + u_t \\ \delta_t &= \delta_{t-1} + v_t \\ \gamma_t &= -\sum_{s=1}^{S-1}\gamma_{t-s} + w_t \end{aligned} \tag{1}$$

The definition of each variable is given below. At time stamp t,

1. y_t is the targeted time series that one wants to predict.
2. $\boldsymbol{x_t}$ is the covariates that one observes along with y_t.

3. μ_t, δ_t are the latent variables of a *local linear trend model*, where both the mean μ_t and the slope δ_t of the trend follow random walks. In particular, $u_t \sim \mathcal{N}(0, \sigma_\mu^2), v_t \sim \mathcal{N}(0, \sigma_\delta^2)$ are i.i.d. noise of the random walks. The intuition of a local linear trend model is that both the mean and slope of the trend component of the time series are free of sudden change as if the trend component has "inertia".

4. γ_t is the latent seasonal variable of a *seasonal model*. The seasonal model can be considered as a regression on S seasonal dummy variables with coefficients constrained to sum to 1 in expectation. The intuition of a seasonal model is that γ_t has periodic values thus tends to encode repeated patterns (seasonal components) of the time series. Additionally, $w_t \sim \mathcal{N}(0, \sigma_\gamma^2)$ is i.i.d. noise.

5. $\boldsymbol{\beta}$ is the weights of a linear regression. Moreover, $\epsilon_t \sim \mathcal{N}(0, \sigma_y^2)$ is i.i.d. noise of the linear regression.

Under the context of Bayesian Inference, we call this model *Bayesian structural time series*(BSTS), and priors are associated with its parameters:

1. $\mu_1, \delta_1, \gamma_1$ follow normal priors.

2. $\sigma_\mu^2, \sigma_\delta^2, \sigma_\gamma^2, \sigma_y^2$ follow inverse Gamma priors.

3. $\boldsymbol{\beta}$ follows *spike and slab prior*[1].

3 Results

3.1 Data Preprocessing

Kaggle (https://www.kaggle.com/datasets/dilwong/flightprices) is the data resource, which collects one-way flight data on Expedia searched between April 16, 2022, and October 05, 2022. It contains 28 variables in total. We removed all flights from Spirit Airline as the data is not reasonable for some covariates.

3.1.1 Bayesian linear regression model

There are 375 missing values in one column and some outliers. We removed these rows with 23574 observations left. After that, we removed the redundant and non-informative columns with 9 variables left. We then tuned *flight departure date* into the weekday (Mon to Sun) and divided the *flight departure time* into four categories (morning, afternoon, evening and night). Besides, we changed the variable named *search date before flight departure date* to a continuous variable – *number of days before flight departure when doing search*.

Figure A.1 shows the exploratory data analysis (EDA) result. The first eight figures plot the relationship between each covariates corresponding to the response variable *flight base fare* and the last figure shows the histogram distribution of the response variable. From EDA, we observed that the flight base fare is heavily dependent on the search days, travel duration and whether the ticket is for basic economy. Particularly, the relationship between flight base fare and search days shows an exponential-decay pattern, prompting a log transformation of this variable. The histogram of flight base fare indicates a normal shape of the variable with a bit of right tail. For simplicity, we transformed the *flight departure days* and *flight equipment type* two variables into binary variables according to EDA.

The final 8 covariates besides the response variable used in linear regression are (with corresponding parameters attached): the number of search days before the flight departure date in log scale ($\beta_{searchDaysBeforeDepartureLog}$); if the flight departs on Tuesday or Wednesday ($\beta_{isFlightTueWed}$); total

[1]In our settings, no regressors (covariates) are provided, therefore we do not model $\boldsymbol{\beta}$ and omit the introduction of the spike and slab prior.

travel duration in hours ($\beta_{travelDuration}$); if the ticket belongs to basic economy ($\beta_{isBasicEconomy}$); the number of seats remaining when searching ($\beta_{seatsRemaining}$); if the flight departs at morning ($\beta_{isDepartureMorning}$); if the airline is Delta ($\beta_{isDelta}$), and if the flight equipment is not Embraer 175 nor Airbus A319 ($\beta_{isEquipmentOther}$). The baseline table summarizing those variables are shown in Table 1. After one-hot encoding on categorical variables, we randomly split the data as training (80%) and test (20%) data for later-on analysis.

Variables	**Median (IQR) or Percent**
Flight base fare ($)	199.1 (134.0, 236.3)
Number of search days before flight departure in log scale	3.26 (2.56, 3.69)
Flight departs on Tuesday or Wednesday (%)	24.1
Total flight travel duration (Hours)	3.0 (3.0, 3.0)
The ticket belongs to basic economy (%)	37.1
Number of seats remaining when searching	9 (7, 9)
Flight departs in the morning (%)	70.4
The airline is Delta (%)	64.7
Flight is not Embraer 175 nor Airbus A319 (%)	81.4

Table 1: Baseline table summarizing the median value for each variable in the data with its interquantile range (IQR) (%25, %75) or percent of data for categorical variables.

3.1.2 BSTs model

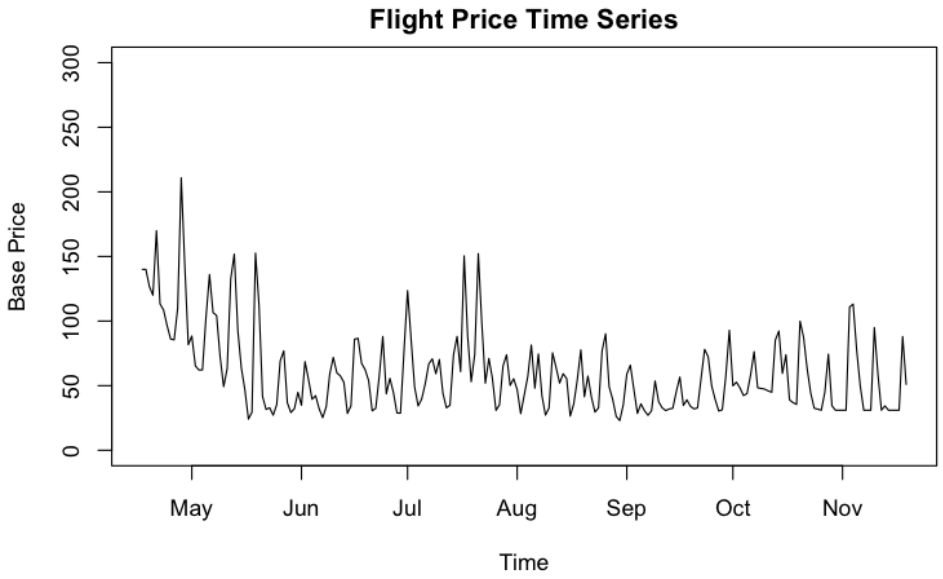

Figure 1: (Average) flight price of DL and AA versus date.

The data preprocessing of the time series version of flight prices are based on the clean data obtained at the previous stage. Specifically, we filter all the non-first-class flight data from Delta and American Airlines. Since there are multiple data points every day, we first aggregate the prices of each day. Concretely, we consider the average flight price as our predictive target. Figure 1 shows the changes in flight prices w.r.t. time. The flight price has a clear periodic pattern (weekly), and is relatively steady after June. We further split the dataset into two partitions: the first 187 days are used for training and the last 30

days are for testing.

3.2 Bayesian Linear Regression Analysis

3.2.1 g-prior

In g-prior linear regression, we set the constant g equal to the sample size of the training data. The inverse-gamma prior parameters for σ^2 were set as $\nu_0 = 1$ and $\sigma_0^2 = SSR(\hat{\boldsymbol{\beta}}_{OLS})/(n-p)$ where $SSR(\hat{\boldsymbol{\beta}}_{OLS})$ is the sum of squared residuals of the ordinary least squares estimate of $\boldsymbol{\beta}$, n is the training data sample size and p is the number of covariates. We ran 5000 iterations in total and the posterior distribution of each parameter is shown in Figure A.2.

Table 2 shows the posterior means and 95 % confidence interval for all parameters. As we can see, each covariate significantly contributes to the flight fare determination. For example, compared to the non-basic economy ticket, the basic economic ticket has a \$76 decrease when controlling other parameters. The Delta airline is typically more expensive than the American Airline by \$30.7 when other variables do not change. Controlling other parameters, the flight departing on Tuesday and Wednesday is \$29.4 cheaper than other days' departure. The parameters we got by g-prior linear regression agree with the EDA result in general.

Parameters	**Posterior Means**	**95% Confidence Interval**
$\beta_{intercept}$	-490.3	(-544.9, -435.5)
$\beta_{travelDuration}$	263.4	(245.8, 281.1)
$\beta_{isBasicEconomy}$	-76.0	(-77.7, -74.2)
$\beta_{seatsRemaining}$	-0.83	(-1.19, -0.46)
$\beta_{isDelta}$	30.7	(28.7, 32.7)
$\beta_{isEquipmentOther}$	-17.6	(-20.0, -15.3)
$\beta_{searchDaysBeforeDepartureLog}$	-28.4	(-29.2, -27.6)
$\beta_{isFlightTueWed}$	-29.4	(-31.1, -27.8)
$\beta_{isDepartureMorning}$	22.3	(20.5, 24.1)
σ^2	2285	(2239, 2332)

Table 2: The g-prior linear regression posterior means and 95% confidence interval for all parameters.

We then did prediction on the test data. For each observation, we sampled 100 points using parameters sampled from the first 100 iterations above and calculated the means as the final prediction. We used RMSE as the evaluation matrix to quantify the quality of prediction. The g-prior linear regression gave us a \$48.1 RMSE, which is not a bad performance. Figure A.3 shows the comparison of the actual flight fare and the g-prior linear regression prediction in test data. We can see that the prediction is better in the lower fare than the higher fare. This makes sense as the number of observations is rare at the higher fare, weighting less in determining the posterior distribution.

3.2.2 Bayesian linear regression model selection

We did Bayesian model selection to further improve the fitting quality. The original g-prior linear regression contains 8 covariates (excluding the intercept) with 5 binary variables and 3 continuous ones. After including all interaction terms and square terms, we got 40 terms in total. As the number of sample size in the training data is too large to run, we randomly sampled 10% observations from them. The rank of the new matrix is 35, indicting a linear correlation between some columns in the matrix. Therefore, we removed 5 related columns (terms), by doing which the new data matrix is invertible.

We applied Markov chain Monte Carlo (MCMC) to figure out all parameters. For $\boldsymbol{\beta}$ and σ^2, we kept the prior the same as g-prior linear regression above. The starting point for $\boldsymbol{z}$ is including all terms in

the model ($\boldsymbol{z_j} = 1 \, \forall j \in [1, p]$). We ran 2000 iterations in total with the effective sample size for $\sigma^2 \approx 1793$ and the average effective sample size for $\boldsymbol{\beta}$ and $\boldsymbol{z}$ is 1043 and 1184, respectively. Figure A.4 shows the posterior probability for parameter $\boldsymbol{z}$. As we can see, there are 14 out of 35 terms make the greatest contribution to the fitting. We then estimated for $\boldsymbol{\beta}$ as the average value from all iterations, called $\hat{\boldsymbol{\beta}}_{bma}$, and applied it on test data prediction. The new RMSE is \$45.3, which is \$2.8 smaller than the g-prior linear regression, indicting a better performance of the model. Figure A.5 shows the comparison of actual flight fare and prediction from model selection result. Same as the g-prior linear regression result, the model selection model also performs worse in the prediction for higher flight fare.

3.3 Bayesian Structural Time Series Analysis

3.3.1 Hyperparameter

There is only one hyperparameter in our model, namely the number of total seasons S. Since we observe a clear weekly repetition of price changes in Figure 1, we set $S = 7$ to capture this pattern.

3.3.2 Priors

In our settings for BSTS model, we only consider the univariate time series, i.e. the average flight prices. There are no other covariates. Therefore, our model does not have any regressor parameters $\boldsymbol{\beta}$. On the other words, we only need to set the priors for $\mu_1, \delta_1, \gamma_1, \sigma_\mu^2, \sigma_\delta^2, \sigma_\gamma^2$. The priors are listed as follow:

1. $\sigma_\mu^2, \sigma_\delta^2, \sigma_\gamma \sim \text{InvGamma}(\nu_0, \nu_0\sigma_0^2)$ where $\nu_0 = 0.01, \sigma_0 = 0.01 \times s_y$. The prior distribution indicates that σ's are small in general without strong confidence. Additionally, we set the upper limit of σ's to be s_y, saying that they cannot be larger than the sample standard deviation of the time series being modeled via `upper.limit` argument of `R` function `SdPrior`.

2. $\mu_1 \sim \mathcal{N}(y_1, s_y^2)$. We set the mean of μ_1 to be the price of the first day in our training set and still using s_y^2 as a reference for variance.

3. $\delta_1 \sim \mathcal{N}(\frac{y_T - y_1}{T-1}, s_y^2)$, where T is the number of data points in the train set. Since δ indicates the slope of linear trend, we use the slope of global trend as a guess for it.

4. $\gamma_1 \sim \mathcal{N}(1/7, s_y^2)$. γ_1 is the component of the first day in a week period, and we can assume the 7 days of a week share nearly equal contribution to the season component.

3.3.3 Training and Prediction

We then fit our model with the training set. We use MCMC to generate 100,000 samples and discard the first 1,000 samples. Figure A.6 shows the decomposition of the time series into two components, seasonality and trend, captured by our model. Furthermore, Figure 2 shows the posterior distribution of μ_t and γ_t. We can conclude the local trend has a increasing tendency, due to the positivity of δ_t. We can expect the flight price to be continuously increasing accordingly. Additionally, in a same week, prices on Monday, Tuesday, and Wednesday are cheaper while prices on Thursday and Friday are significantly more expensive. To our surprise, the flight prices at the weekends are not higher than the average price.

We then use the trained model to predict the flight prices of the next 30 days. Figure 3 shows the prediction of our BSTS model. Compared with Figure 1, our model captures the weekly pattern and has a slightly increasing trend, and has a in-sample RMSE of 20.84 using the mean of predictive distribution, showing it fits the data well. However, the RMSE of the prediction is 25.34, not significantly lower than the standard deviation of the testing set 26.27, showing our fitted model does not predict the out-of-sample distribution well.

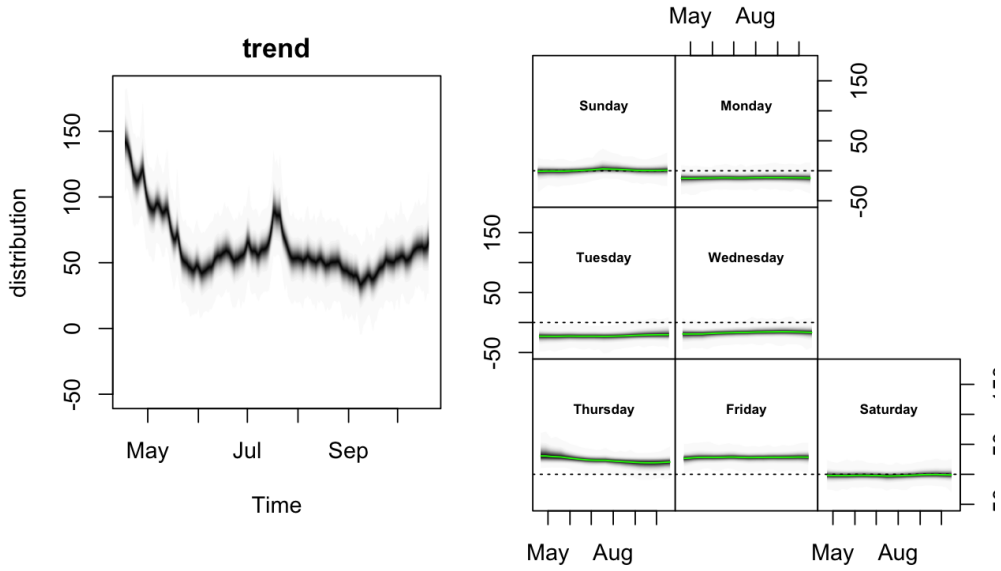

Figure 2: Distribution of two components. The left plot shows the posterior distribution of μ_t. The right plot shows the posterior distribution of γ_t.

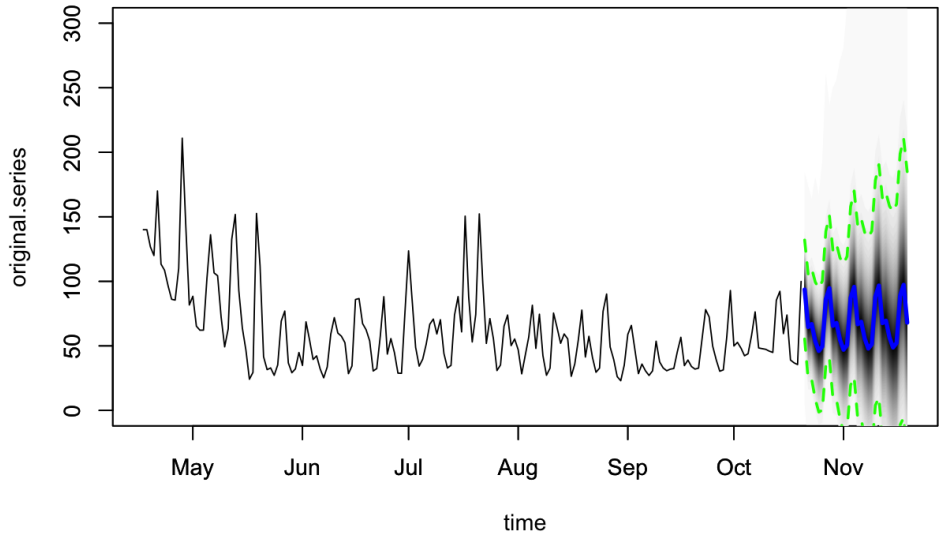

Figure 3: Prediction of BSTS. The blue line is the mean of the predictive distribution. The outer green dash lines are the 95% confidence intervals of the prediction.

4 Conclusion

4.1 Major Findings

In this project, we successfully made prediction of flight fare through Bayesian linear regression and Bayesian structural time series analysis. These two models focus on different parts of the data and can not be compared directly. The g-prior linear regression produced the posterior for all parameters and the confidence interval indicated all 8 covariates playing a significant part in predicting flight fare. The

RMSE is $48.1 with an underestimation in higher flight fare prediction, probably due to a rare data density in that range. The model selection included interaction and square terms into the model and picked out 14 important terms by which the RMSE decreased to $45.3.

In the time series analysis through the BSTS model, we successfully decomposed the mean daily flight price into two components, trend, and seasonality. Our fitted model resulted in a RMSE of $20.84 in the training set but has a significantly higher RMSE of $25.34 on the testing set, showing our model failed to generalize well out of the training distribution. However, our trained model still yield some valuable posterior information: we found prices are higher on Thursday and Friday while cheaper from Monday to Wednesday in the same week, aligning with the discovery from the linear regression model. A list of potential failure reasons is given in the next section.

4.2 Limitation

There are some limitations in our project. For the data, some values are not reasonable so that we have to remove them. It is necessary to dig into the source of the data and figure out the problem. The g-prior linear regression performs bad in higher flight fare prediction, where one possible reason is the right tail in the distribution. We could try t-distribution instead and compare the result. Also, if given more time, we will use all training data in model selection.

As for the BSTS model, the RMSE we get is not largely lower than the standard deviation of the test data, which shows that our model does not perform well on out-of-sample data. We think the error might come from these aspects:

1. The length of the time series is too short to capture the yearly pattern of the flight price. In particular, we only have flight price data for 217 days, and it is technically impossible to model the seasonal component for a year.

2. The data collection process contains unreported bias. The dataset we used in this report has not been widely used by other works since it was released a few months ago. During our EDA, we also find the datasets might contain systematic bias. For example, the number of remaining seats is missing for all Spirits' flights.

3. The univariate time series are of low representative power. In other words, it is fundamentally difficult to model the time series of the collected flight price given even an infinite number of data. This may be because many latent variables that influence the final mean daily price are unobserved.

References

[1] Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*, 1986.

[2] Steven L. Scott and Hal R. Varian. Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2):4–23, 2014. doi: 10.1504/IJMMNO.2014.059942. URL https://www.inderscienceonline.com/doi/abs/10.1504/IJMMNO.2014.059942. PMID: 59942.

A Appendix

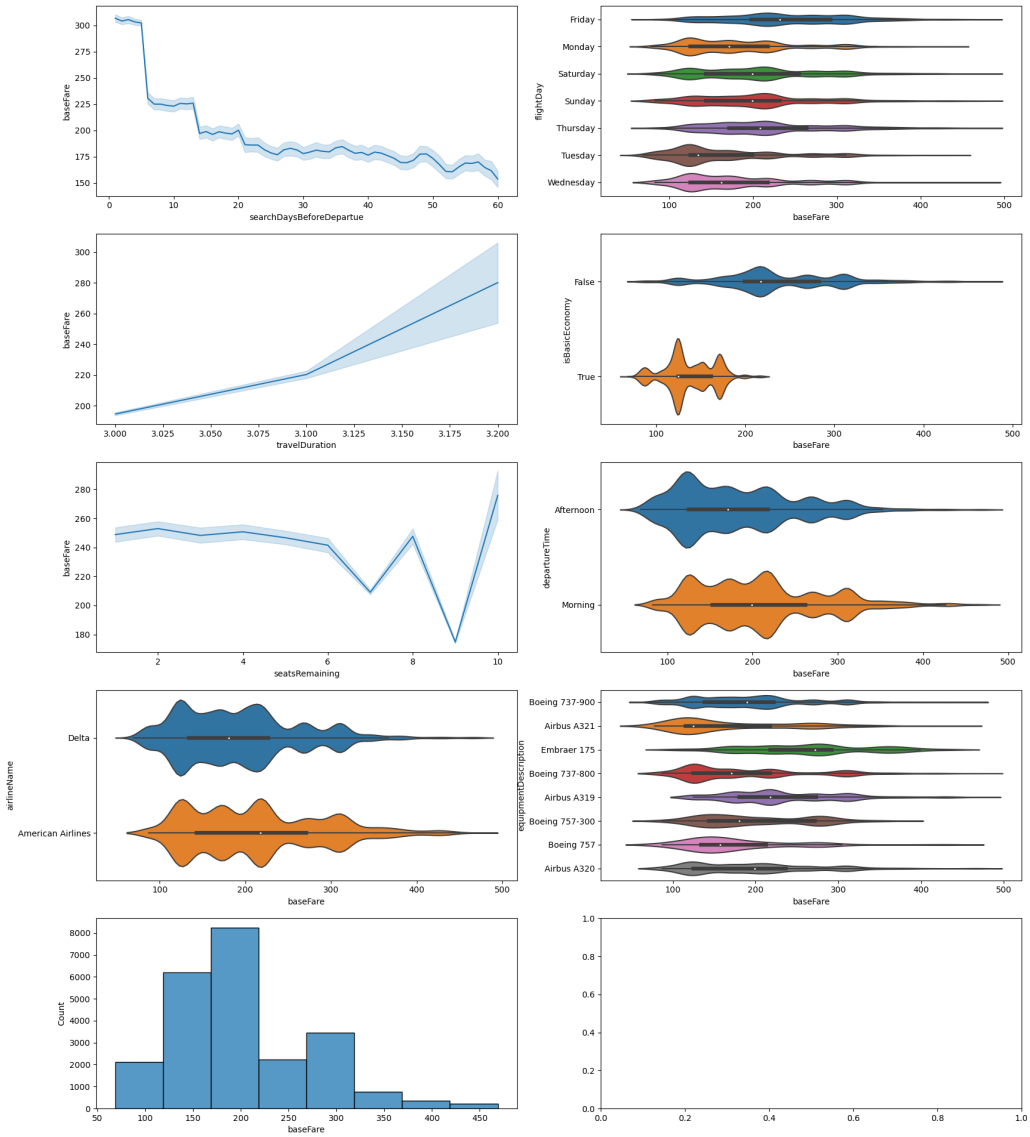

Figure A.1: The exploratory data analysis plot of all variables corresponding the response variable flight fare.

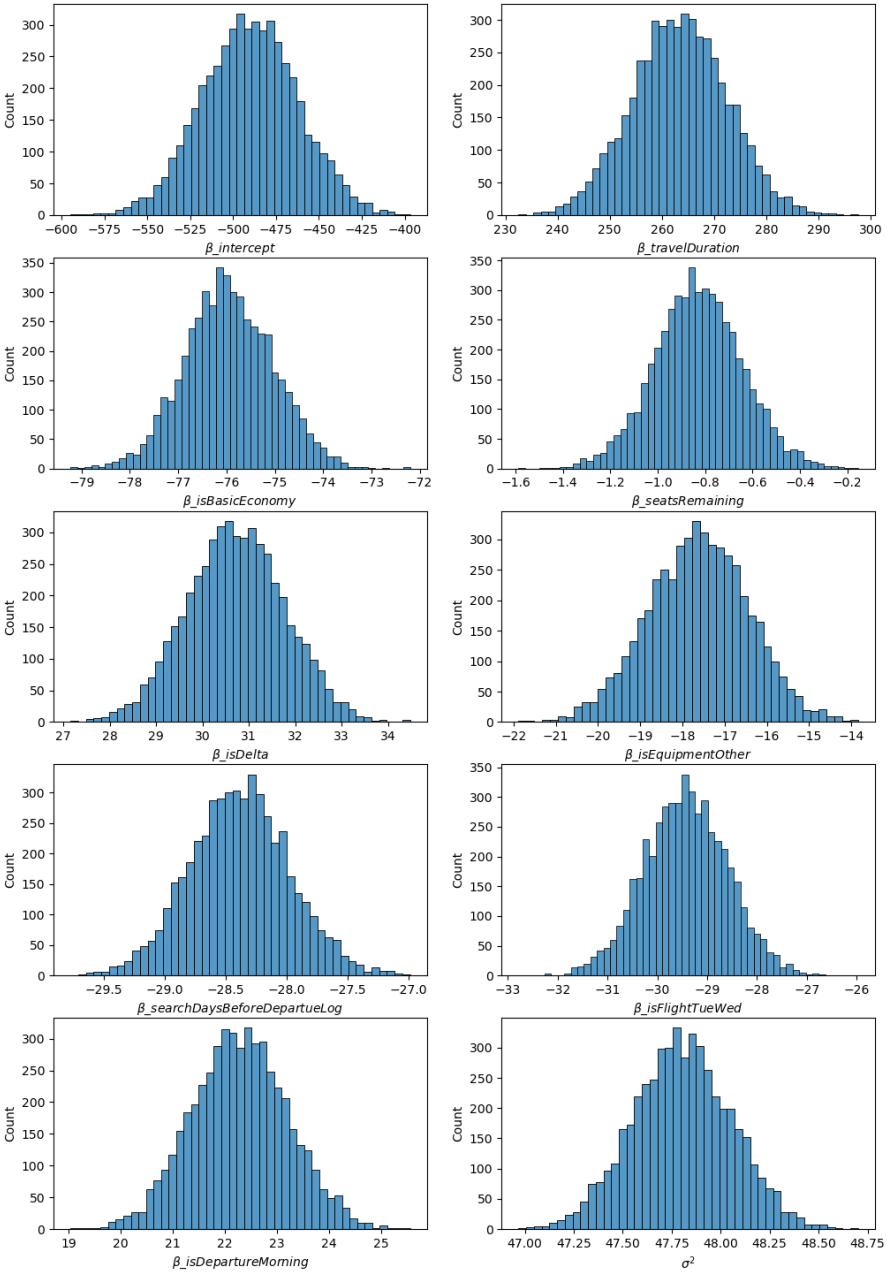

Figure A.2: The g-prior linear regression posterior distribution for all parameters.

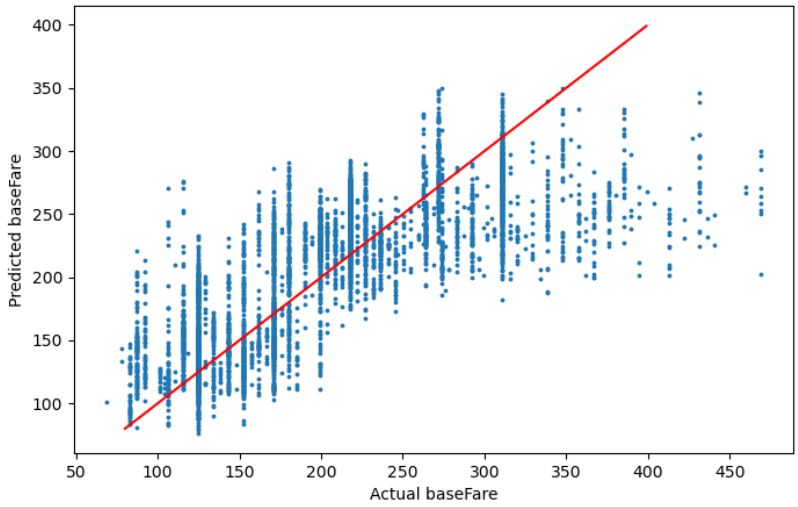

Figure A.3: The comparison between the actual flight fare and g-prior prediction in test data.

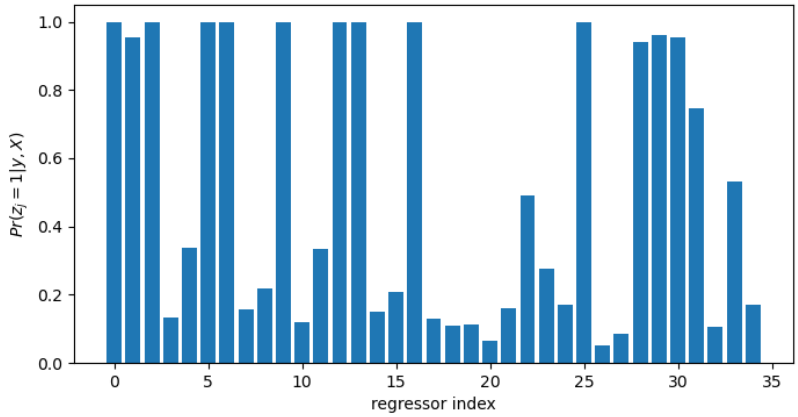

Figure A.4: The posterior probability for $\boldsymbol{z}$.

B Contribution

Zihan Cui Zihan Cui focused on construction of Bayesian models and model selection. He wrote the introduction of the project in report, showed methods used and made a summary of conclusions. He also made slides and participated in recording.

Yushan Yang Yushan worked on the data preprocessing and Bayesian linear regression part of the project. She wrote all codes, wrote paragraphs in report and made slides for presentation related to those parts.

Xingjian Zhang Xingjian is in charge of the BSTS part of the project. Specifically, he did the data cleaning of the time series, wrote all the R codes about BSTS modeling, wrote the BSTS part in the report, and covered the corresponding presentation.

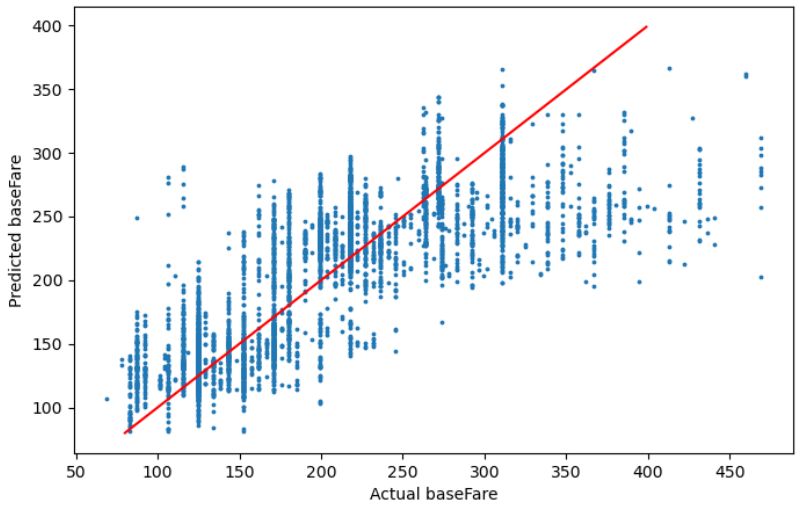

Figure A.5: The comparison between the actual flight fare and model selection prediction in test data.

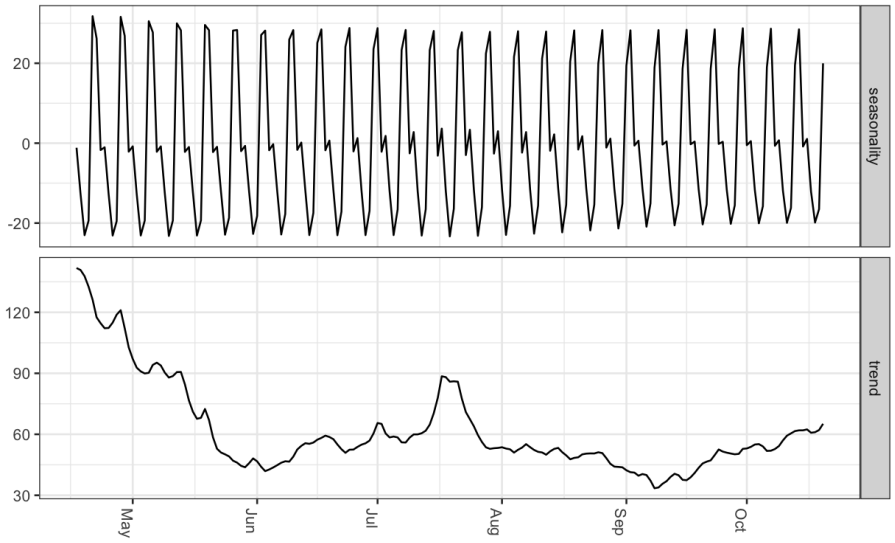

Figure A.6: The decomposition of two components: Seasonality and Trend.