Bayesian Logistic Prediction of Air Flight On-Time Performance in the United States

Yifan Zhu Student No.1006345849

21/12/2020

Abstract

Airline flight delays have led to a great deal of interest in understanding flight on-time performance. Flight delays bring additional costs and inconvenience to both travellers and airline carriers, and are often attributed to several causes, including weather conditions, airport congestion, airspace congestion (Deshpande & Arıkan, 2012). In this paper, we analyze the empirical commercial flight data, published by the U.S Bureau of Transportation Statistics of the 17 major airlines flying within the United States in December 2019. This study will investigate the schedule adherence and implement a multilevel Bayesian logistic regression. We will arrive at a statistical model which can forecast the on-time departure probability of commercial flights by airline and by day of the week. These model predictions will have implications for passengers, airport regulators and airline managers.

Key words: flight delays, Bayesian, logistic regression, statistical modelling, forecasting

Repository: https://github.com/YvonneYifanZhu/STA304FinalProject_YifanZhu.git

Introduction

On-time performance, or schedule adherence, refers to the level of success of a transportation service (such as buses or trains) adhering to the published schedule. For many transport systems, the level of on-time performance is an important measurement of the effectiveness of the system ("On-time performance", 2020). Flight delays have had a large and far-reaching impact on the U.S. economy for many years. The cost of flight delays includes the operating costs of airlines, passenger delay costs, and jet fuel costs, as well as the additional carbon dioxide disruption. According to a report published by the Joint Economic Committee of the U.S. Congress, the cost of domestic air traffic delays to the U.S. economy was as much as $41 billion for 2007. The report notes this cost is particularly high in December, as the height of holiday traveling resulting in, corresponding flight delays that add up to almost 438,000 hours.

According to the U.S. Department of Transportation (DOT), a flight is considered delayed if its arrival (or departure) time at the gate is 15 minutes or more after the scheduled time. The difference between actual departure time and scheduled time is quite uncertain for many reasons, as those mentioned above. Many of these reasons are hard to foresee when passengers make their purchasing decisions. But unless they buy travel insurance when they book the flights, delayed U.S. passengers are out of luck. Under current EU legislation, passengers traveling from within the EU with an EU-based carrier, have the obligation for any compensation if the delayed schedule exceeds three hours; however, in the United States, there are no federal laws asking airlines to compensate passengers with money or offer other means of compensation when the flight is delayed. It is entirely up to airlines, only some U.S. airlines such as American Airlines offers similar compensation.

Potential flight delays and their underlying costs not only affect passengers making purchase decisions, but also airline carriers making scheduling decisions. The aim of this study is to investigate the on-time performance and construct a Bayesian logistic model for predicting the on-time departure probability of a flight. We will examine whether flight delay probability differs by the day of the week, by airline carrier, and the length of the flight.

Our research questions have implications for airport regulators, airline managers and passengers. For airport regulators, our findings can help them prepare for stranded passengers and a backlog of flights in advance. In the ideal case, airline managers can make better scheduling decisions to minimize operational costs. By comparison with their competitors in the flight industry, carriers can make business responses to improve their management. Finally, if passengers adopt our prediction results, they will have a tool to evaluate the on-time probability of their flights when purchasing tickets and be better able to choose a flight, and decide if it is worth buying the travel insurance.

Methodology

Data

The U.S. Bureau of Transportation Statistics (BTS) is part of the U.S. Department of Transportation (DOT). The credible content provided by BTS helps decision makers and the public better understand statistics on transportation. The data used in this study is the Airline On-Time Performance data, which collects all scheduled domestic flights by major commercial flight carriers in the United States since 1987(updated by month), which can be directly downloaded from the Bureau of Transportation Statistics website. The data cleaning process can be found in the section "Appendix" or through online repository (https://github.com/YvonneYifanZhu/STA304FinalProject_YifanZhu.git).

The target population of this study is all domestic commercial flights in the United States; the sampling frame in this paper is the flights scheduled in December 2019. As mentioned above, December is the peak time of travel, so the data is comprehensive, representing the full-load transportation system. In addition,we chose the December 2019 dataset to reduce the impact of COVID-19 on the results, as this is before many countries went into lockdown, and airlines canceled or rescheduled many flights. Thus, this sampling frame should be good to represent the target population, which contains 618612 observations with 8 selected variables:

- DayofMonth (31 levels): days of the month
- DayofWeek (7 levels): days of the week
- Cancelled (2 levels): cancelled flight indicator (1= Yes, 0=No)
- Carrier (17 levels): unique 2-letter carrier code
- DepDelay (discrete): the difference between the scheduled and actual departure time in minutes; early departure is assigned to negative numbers.
- DepDel15 (2 levels): departure delay indicator if 15 minutes or more (1= Yes)
- AirTime (discrete): flight time in minutes

Data Preparation

The non-response data has been named NA in the data frame. In order to process this large data set more efficiently in RStudio during modeling, the author eliminated some missing data and randomly sampled 6000 observations from the sampled population of 618612 observations. In the following study, we will only use the 6000 newly sampled observations and build a model using DayofWeek, Airline, AirTime to predict the delay probability of the flight. See the appendix or GitHub repository link for details on how the data set was cleaned and sampled.

Table 1: Sampled Data Set

DAY_OF_MONTH	DAY_OF_WEEK	OP_UNIQUE_CARRIER	DEP_DELAY	DEP_DEL15	CANCELLED	AIR_TIME
1	7	HA	-1	0	0	325
1	7	HA	-2	0	0	287
1	7	HA	5	0	0	323
1	7	HA	142	1	0	289
1	7	HA	1	0	0	359
1	7	HA	10	0	0	299

Data Visualization

This section briefly explores the raw data containing 618612 flights during December 2019.

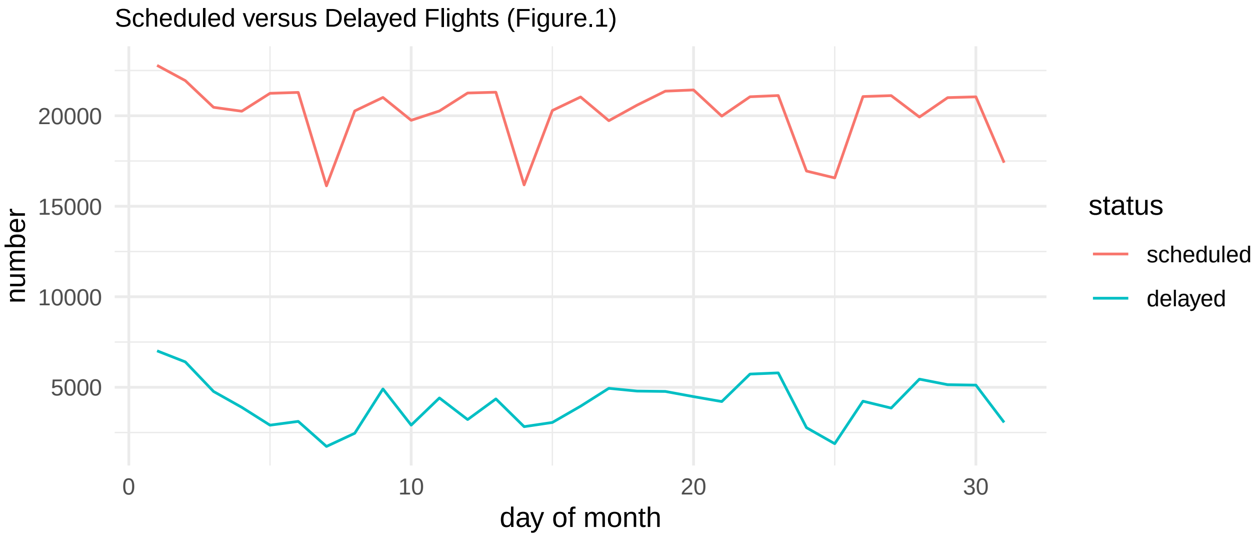


Figure.1: This multiple-line plot compares the number of scheduled flights and the delayed flights during December. We can find the total number of flights has a fluctuating weekly trend, along with three dramatic drops in numbers. As for the delayed flights, there is no obvious pattern over one month, but it seems to have a correlation with the total number of flights. We compute the correlation coefficient in R, which is 0.535 (positive correlation). Therefore, we can conclude that the delayed probability associates with the total number of scheduled flights.

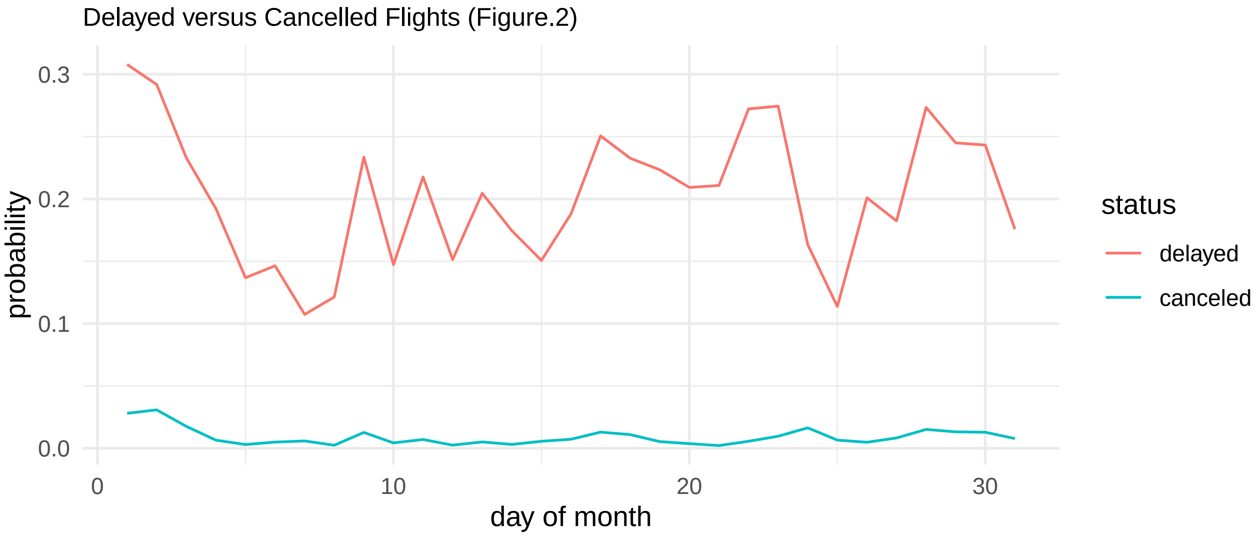


Figure.2: The delayed and canceled probabilities are combined in this figure. It is surprising to find that while passengers face the different probabilities of delays, the probabilities of canceled flights remain stable over the month. Even when the delayed number is quite high, there is only a slight change in canceled flights. We also compute the correlation coefficient between total flights and canceled flights (0.234), which is smaller than the previous one (matches with our finding).

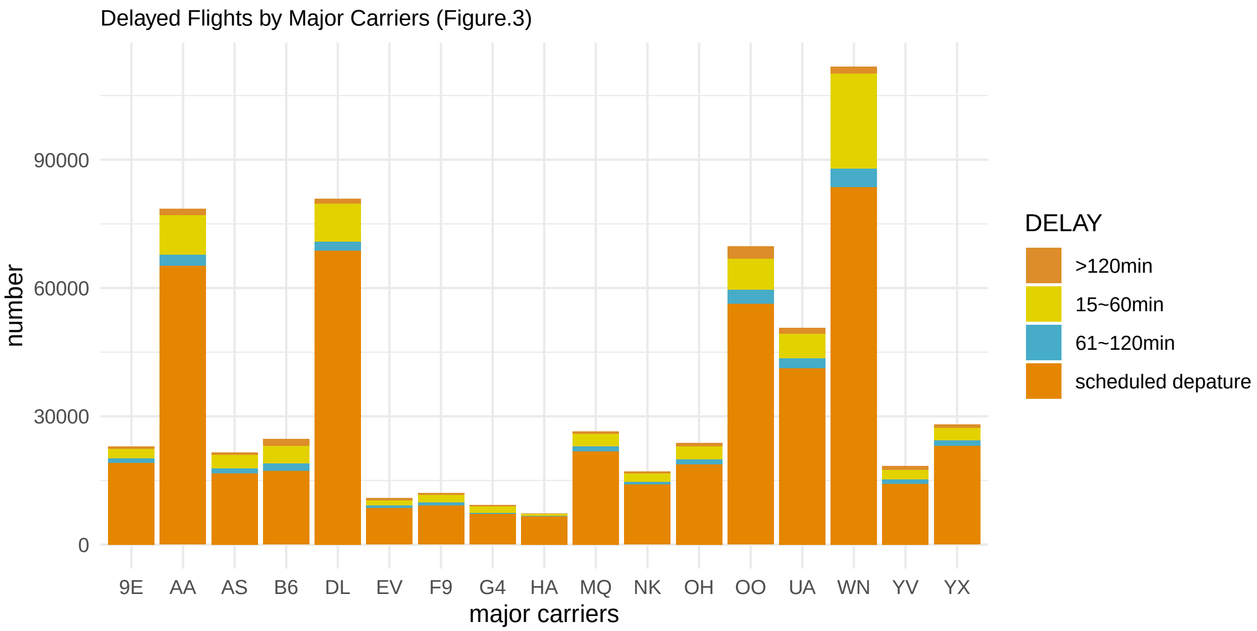


Figure.3: This stacked bar plot shows the on-time performance of each airline. Apart from canceled flights, Southwest Airlines (WN), Delta Air Lines (DL), American Airlines (AA), Sky West Airlines (OO), and United Airlines (UA) are the most active carriers during December. We can also find that Southwest Airlines owns the dominant position compared to other airlines. According to BTS, Southwest carried the most domestic passengers of any United States airline in 2018, which matches our results in December 2019. However, Southwest Airlines is also considered to have the most delayed flights. Moreover, we find that among the delayed flights, most of the flight delays within 1 hour and there is only a small proportion of flights whose departure time is 2 hours later than the scheduled departure time.

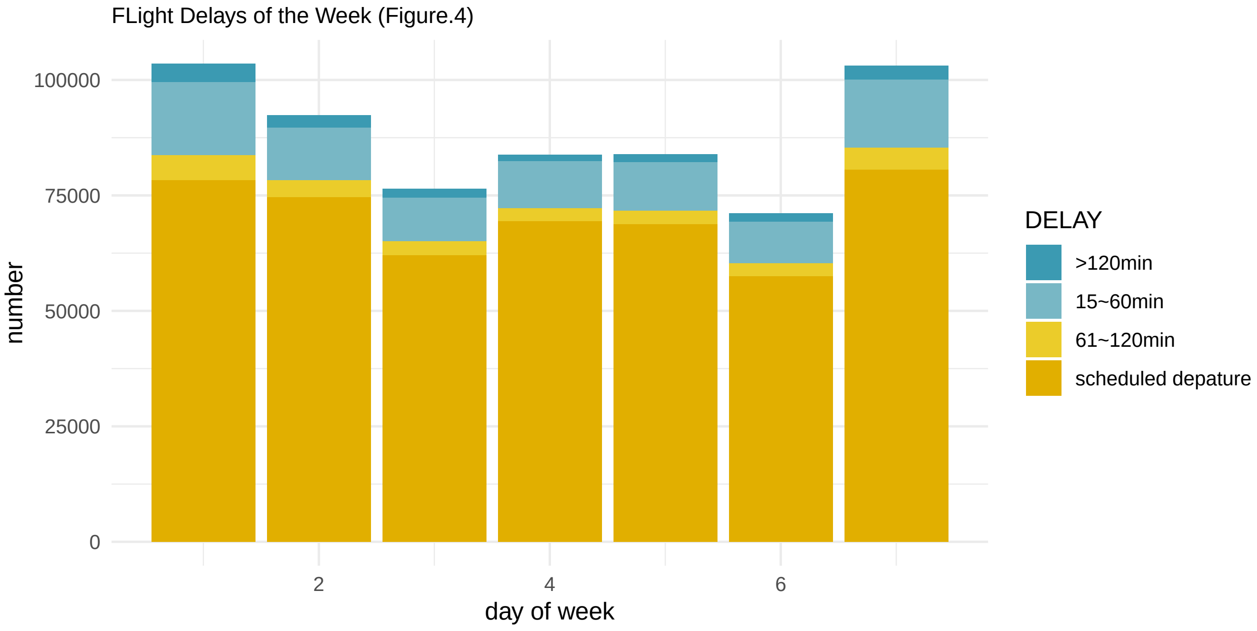


Figure.4: The stacked bar chart visualizes the delayed time over the week. The transportation system is the busier on Monday and Sunday, while Wednesday and Thursday do not experience the high volume of traffic. Although most of the flights during the week delay within 1 hour, Monday has the highest number of flights that depart 2 hours after the scheduled time.

Model

Model Selection

A Bayesian logistic model is employed to address this study. Logistic regression is used to model the probability of a binary variable with only two potential outcomes labeled “0” and “1” while the predictor variables are continuous or categorical. There are several assumptions of the logistic regression model: First, binary regression requires the outcome variable to be binary; Second, the observation should be independent of each other. Third, there should be little or no multicollinearity among predictors; Fourth, independent variables have linear relationship with log odds; Fifth, large sample size is necessary to carry out the logistic model. The natural logarithm of an odd ratio, logit, is the central mathematical concept of the logistic regression. (Peng, Lee, & Ingersoll, 2002). The logit can be stated as:

$$logit(P_i) = ln(P_i/(1-P_i)) = \beta_0 + \beta_1 X_i$$

Convert odds to simply probability:

$$P = \frac{e^{\beta_0+\beta X}}{e^{1+\beta_0+\beta X}}$$

Why Bayesian? Bayesian models use multiple and simultaneously running Markow chains that adopt priors to inform posteriors to update new posteriors until all the chains converge. Bayesian inference for logistic analyses follows the usual pattern of Bayesian analyses: 1) Write the likelihood function. Since the delay outcome is binary, $Likelihood = P^Y(1-P)^{(1-Y)}$, $Y = 1$ if delay, otherwise equals 0; 2) Form a prior distribution over all unknown parameters, the overall likelihood is the summation of all flights in the data set. The most common prior for logistic regression parameters is $\beta \sim N(\mu_0, \sigma_0^2)$; 3) Find the posterior distribution over all parameters. The posterior distribution is derived by multiplying the prior distribution over all parameters by the full likelihood function (Kana, 2020). Therefore, the Bayesian logistic regression model is given by

$$P(\beta \mid data) \propto P(data \mid \beta)P(\beta), \beta \sim N(\mu_0, \sigma_0^2)$$

$$P(\beta_0, \beta_1, ..., \beta_k) \propto P(y, X_1, ..., X_k \mid \beta_0, \beta_1, ..., \beta_k)P(\beta_0, \beta_1, ..., \beta_k)$$

The brms package in R offers a wide range of Bayesian single-level and multilevel models. Apart from brms package, there are many other packages developed to fit MLMs in R, which are limited insofar since they are only predicting the mean of the response distribution and their assumptions may be violated in many applications such as constant residual standard deviation in linear models. (Bürkner, 2018). An alternative model in the study can be a ordinary logistic model using package lme4 in R. However, because of the intensive nature and build-in flexibility of Bayesian modeling, Bayesian models end up being more precise. Therefore, we choose to apply the Bayesian single-level regression model in the following analysis. Assume the flight is schedule at day i and airport j, P stands for the probability of flight delay. Furthermore, $i = 1, 2, ..., 7$ represents the day of the week, from Monday to Sunday respectively. Since there are 17 major carriers in this study $j = 1, 2..., 17$, it follows the order 9E, AA, AS, B6, DL, EV, F9, G4, HA, MQ, NK, OH, OO, UA, WN, YV, YX. The natural log of odds:

$$logit(P) = ln(P/(1-P)) = \beta_0 + \beta_1 X_{AirTime,ij} + \sum \beta_i X_{DayofWeek,i} + \sum \beta_j X_{Airline,j} + \epsilon$$

- $X_{AirTime_i j}$: If passenger purchases the flight on day i, then $X_{AirTime_i} = 1$ and $X_{AirTime,\neq i} = 0$.
- $X_{DayofWeek,i}$: since there are 7 days of the week,$i = 1, 2..., 7$.
- β_i: the coefficient of each day of the week.
- $X_{Airline,j}$: If passenger purchases the flight by airline j, then $X_{Airline_j} = 1$ and $X_{Airline,\neq j} = 0$.
- β_j: $j = 1, 2..., 17$, the coefficient of variable $X_{Airline,j}$
- Convert odds to simply probability

$$P = \frac{e^{\beta_0+\beta_1 X_{AirTime_i j}+\sum \beta_i X_{DayofWeek,i}+\sum \beta_j X_{Airline,j}+\epsilon}}{e^{1+\beta_0+\beta_1 X_{AirTime_i j}+\sum \beta_i X_{DayofWeek,i}+\sum \beta_j X_{Airline,j}+\epsilon}}, i = 1, 2...7, j = 1, 2, ...17$$

Model Diagonstics

Similar to logistic model, there are five assumptions that need to be investigated: correct specification of the model, linearity, independence, equal variance, and normality. There are many tools to check the Bayesian regression model. Apart from the results shown in the appendix, we can use the package(shinystan) which can directly produce the diagnostic plots. Figure.5: Posterior Predictive Graph. Posterior predictive checkBased on the graphical check, there is no significant systematic discrepancies of our data from what can be predicted from our model.Hence, we have a correct specification of this Bayesian model; Figure.6: Non-convergence Plot. We use the mcmc_trace function from the bayesplot package to plot the caterpillar plot for each parameter. From the series of plot below, four chains mix well.

Modeling Results

Table 2: Population-Level Effects

	estimated.coefficients	lower.95.CI	upper.95.CI
Intercept	-1.60	-1.97	-1.23
Tuesday	-0.24	-0.47	-0.02
Wednesday	-0.25	-0.48	-0.02
Thursday	-0.49	-0.72	-0.25
Friday	-0.21	-0.43	0.01
Saturday	-0.21	-0.45	0.03
Sunday	-0.05	-0.27	0.16
American Airlines (AA)	0.10	-0.30	0.49
Alaska Airlines (AS)	0.34	-0.14	0.82
JetBlue Airways (B6)	0.95	0.52	1.40
Delta Air Lines (DL)	-0.09	-0.49	0.33
Atlantic Southeast Airlines (EV)	0.50	-0.06	1.10
Frontier Airlines (F9)	0.72	0.16	1.28
Allegiant Air (G4)	0.72	0.18	1.25
Hawaiian Airlines (HA)	-0.46	-1.51	0.44
Envoy Air (MQ)	0.32	-0.13	0.77
Spirit Airlines (NK)	-0.08	-0.69	0.48
Comair (OH)	0.34	-0.13	0.83
Sky West Airlines (OO)	0.15	-0.24	0.56
United Airlines (UA)	0.23	-0.17	0.65
Southwest Airlines (WN)	0.64	0.27	1.03
Mesa Airlines (YV)	0.27	-0.26	0.78
Repblic Airways (YX)	0.19	-0.27	0.64
AirTime	0.00	0.00	0.00

Table.2: This table contains the regression results of the Bayesian logistic model. The respective coefficients are listed in the cells. Since P=1 if flight delays, the higher estimated coefficient means it is more likely to delay. Compared to Monday, all the coefficients are negative; thus, Monday has the highest probability to experience flight delays while Thursday has the lowest probability on average.

As for the airlines, respective to Endeavor Air (9E), Delta Air Lines (DL), Hawaiian Airlines (HA), Spirit Airlines (NK) have better on-time performance. Since JetBlue Airways's coefficient is the largest (0.95), the flights scheduled by JetBlue is the most likely to delay. On the other hand, Hawaiian Airlines (-0.46) is predicted to have the highest on-time rate compared to other major carriers. Also, as discussed above, Southwest Airlines has the highest number of total delayed flights but according to our results, Southwest Airlines does not have a higher probability to delay.

The total number of delay is mainly contributed by the number of scheduled flights. Therefore, it may be one of the reasons that most domestic passengers in the United States choosing Southwest Airlines.

Longer flights always come with more retributions and regulations that airlines need to deal with. In addition, since their routes involved multiple airspaces and airlines need to corporate with more conditions (such as the changing weather along the route). However, it is interesting to find that the air time does not show any significant impact on flight delays. To minimize the potential cost, the best case is purchasing a flight that plans on Thursday carried by Hawaiian Airlines. Try to avoid the worst scenario: booking a flight that is scheduled on Monday carried by JetBlue Airways.

Discussion

Model Summary

From the model results, we find that flight delay probabilities vary across the day of the week and different airline carriers, but it is surprising to find that flight time has no significant contribution to determine flight delays. Since CIs for most of the variables in the population level do not contain zero (means p-value <0.05), there is strong statistical evidence to illustrate that flight delays are highly associated with day of the week and the selected carrier. As the estimated coefficients shown in the table, we can compute the odd ratio for each category; thus, we can access the delay probabilities of all flights in general. To be more precise, Monday has the highest probability (P=0.172) to encounter flight delays while Thursday is the lowest (p=0.110). Also, choosing airlines is essential to flight delays. In our results, purchasing a flight ticket of JetBlue Airways (B6) is the most likely to delay while Delta Airlines has the best on-time performance on average.

Conclusion

This paper presents a single-level analysis for the on-time performance of commercial flights in the United States. Our results indicate that the on-time departure probability depends significantly on the day of the week and airline carriers. According to the estimation, Monday has the highest probability to delay at the population-level for all airlines. One hypothesis is that airlines and airports are experiencing the highest volume of flights on Monday, making it harder to dispatch aircrafts. For passengers, it can be a little more strategic on buying flight tickets. To minimize your chances of delays, book your flights on Wednesday and Thursday or choose the carriers such as Delta Airlines, Hawaiian Airlines, and Spirit Airlines if possible. If you cannot avoid traveling on Monday, purchasing travel insurance is worth considering.

For airports dealing with a foreseeable rising number of flight delays, in addition to expanding and modernizing airport capacity in the long-term, airports should increase efficiency of existing resources by upgrading utilization and updating the air traffic management system. The results also give implications to airline managers as comparing the on-time performance within the industry. Except looking for the delay probability, they should investigate the structure of the airline's flights. For example, if this carrier has more scheduled lines depart/ arrive at the busier commercial airports (e.g., Hartsfield-Jackson Atlanta International Airport, Los Angeles International Airport, John F. Kennedy International Airport), it is more likely to raise the prediction to flight delays on average.

Weakness & next steps

The work might have some drawbacks as follows. Firstly, our research might have omitted variable bias problems. Apart from the day of the week, airline carrier, airtime, there are many other variables that potentially have a significant impact on flight delays, such as the scheduled time, departure/ arrival airports, the passenager capacity of a flight, etc. By omitting these variables, existing variables' estimated coefficients and significance may be biased; hence, leading to an invalid inference. Secondly, logistic regression and Bayesian model require the observations to be independent of each other which is not satisfied in this study. This is because one flight delay will always influence the following flights. Nevertheless, it is still possible to build a model from dependent observations (Daskalakis et al., 2019). Lastly, since the data in this study collects the information of December only, the evidence is not enough to arrive at a more generalized conclusion. Although the transportation system in December is the closest to its capacity, if we want to study the delay probability on a yearly basis, we should conclude the data from January to December.

Related to the hypothesis that flight delays may perform differently within departure (or arrival) airports, a further study is therefore suggested to investigate the group-level effect within airports using a multilevel regression model. By adding a layer in the model, respective coefficients at the population-level and group-level can show whether the three explanatory variables have the same effect across all airports. Furthermore, we can focus on one airport or simply just one flight. By focusing on one flight, flight carrier, departure/ arrival airport, and flight time are fixed; thus, using the historical data, we can arrive at a more accurate model to predict the on-time performance of this flight by day of the week.

Reference

On-Time : Reporting Carrier On-Time Performance (1987-present). (n.d.). Bureau of Transportation Statistics. Retrieved December 20, 2020, from https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

Peng, C., Lee, K., & Ingersoll, G. (2002). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research, 96(1), 3-14. Retrieved December 21, 2020, from http://www.jstor.org/stable/27542407

Bürkner, P. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395. doi:10.32614/rj-2018-017

Deshpande, V., & Arıkan, M. (2012). The Impact of Airline Flight Schedules on Flight Delays. Manufacturing & Service Operations Management,14(3), 355-484. Retrieved May 4, 2012, from https://pubsonline.informs.org/doi/abs/10.1287/msom.1120.0379.

On-time performance. (2020, April 24). Retrieved December 12, 2020, from https://en.wikipedia.org/wiki/On-time_performance

THE JOINT ECONOMIC COMMITTEE (2008). Your Flight Has Been Delayed Again: Flight Delays Cost Passenger , Airlines and the U.S. Economy Billions https://www.jec.senate.gov/public/_cache/files/47e8d8a7-661d-4e6b-ae72-0f1831dd1207/yourflighthasbeendelayed0.pdf

Kana, M. (2020, February 21). Introduction to Bayesian Logistic Regression. Retrieved from https://towardsdatascience.com/introduction-to-bayesian-logistic-regression-7e39a0bae691

Daskalakis, C., Dikkala, N., & Panageas, I. (2019, May). Regression from Dependent Observations. http://people.csail.mit.edu/costis/dependent-regression.pdf

Yu, R., & Abdel-Aty, M. (2013). Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes. Accident Analysis & Prevention, 58, 97-105. doi:10.1016/j.aap.2013.04.025

Software used in producing the report: Rstudio

Packages used in producing the report:

dbplyr: A Grammar of Data Manipulation

kableExtra: Construct Complex Table with 'kable' and Pipe Syntax

ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics

brms: Bayesian Regression Models using 'Stan'

reshape2:Flexibly Reshape Data: A Reboot of the Reshape Package

plyr: Tools for Splitting, Applying and Combining Data

latticeExtra: Extra Graphical Utilities Based on Lattice

wesanderson: A Wes Anderson Palette Generator bayesplot: extensive library of plotting functions for use after fitting Bayesian models

Retrieved December 20, 2020, from https://cran.r-project.org/web/packages/available_packages_by_name.html

Appendix

Repository: https://github.com/YvonneYifanZhu/STA304FinalProject_YifanZhu.git

Set up and import data

```
# set up
library(tidyverse)
library(dbplyr)
library(brms)
library(ggplot2)
library(plyr)
library(reshape2)
# install.packages("lattixeExtra")
library(latticeExtra)
library(wesanderson)

# download from Bureau
# https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time
airline <- read.csv("/Users/yvonne_zhu/Desktop/STA304 Final Project_Dec.21/flights.csv")
airline <- airline[,1:(length(airline)-1)]
```

```
# data cleaning
# divide delay minutes into categories
airline <- airline%>%
mutate(DELAY = case_when(
    DEP_DELAY <15 ~ "scheduled depature",
    DEP_DELAY >15 & DEP_DELAY <=60  ~ "15~60min",
    DEP_DELAY >60 & DEP_DELAY <=120  ~ "61~120min",
    DEP_DELAY >120  ~ ">120min"
  ))
```

```
# group flights for each day - data visualization
# groupby each day

# total, delayed, canceled flights
total <- count(airline,"DAY_OF_MONTH")
date_delay <- aggregate(DEP_DEL15~DAY_OF_MONTH, data=airline, FUN=sum)
date_cancelled <- aggregate(CANCELLED~DAY_OF_MONTH,data=airline, FUN=sum)

joined_df <- merge(total, date_delay,by="DAY_OF_MONTH")
merged_date <- merge(joined_df, date_cancelled,by="DAY_OF_MONTH")

merged_date$prob1 <- merged_date$DEP_DEL15/merged_date$freq
merged_date$prob2 <- merged_date$CANCELLED/merged_date$freq
```

```r
# subset for carriers' total/delay/cancelled flights
carrier <- count(airline,"OP_UNIQUE_CARRIER")
carrier_delay <-aggregate(DEP_DEL15~OP_UNIQUE_CARRIER, data=airline, FUN=sum)
carrier_cancelled <- aggregate(CANCELLED~OP_UNIQUE_CARRIER,data=airline, FUN=sum)

joined_df2 <- merge(carrier,carrier_delay,by="OP_UNIQUE_CARRIER")
merged_carrier<- merge(joined_df2,carrier_cancelled,by="OP_UNIQUE_CARRIER")

merged_carrier$prob1 <- merged_carrier$DEP_DEL15/merged_carrier$freq
merged_carrier$prob2 <- merged_carrier$CANCELLED/merged_carrier$freq
```

```r
# total, delay, canceled by date of the week
week <- count(airline,"DAY_OF_WEEK")
week_delay <- aggregate(DEP_DEL15~DAY_OF_WEEK, data=airline, FUN=sum)
week_cancelled <- aggregate(CANCELLED~DAY_OF_WEEK, data=airline, FUN=sum)
merged_week <- merge(week,week_delay, by="DAY_OF_WEEK")
merged_week <- merge(merged_week,week_cancelled, by="DAY_OF_WEEK")

merged_week$prob1 <- merged_week$DEP_DEL15/merged_week$freq
merged_week$prob2 <- merged_week$CANCELLED/merged_week$freq
```

Data Visualization

Figure.1 Scheduled versus Delayed Flights

```r
df_plot1 <- select(merged_date, DAY_OF_MONTH, freq, DEP_DEL15)
colnames(df_plot1)<-c("DAY_OF_MONTH","scheduled","delayed")
plot1 <- melt(df_plot1,  id.vars = 'DAY_OF_MONTH', variable.name = 'status')
ggplot(plot1, aes(DAY_OF_MONTH, value)) +
  geom_line(aes(colour = status))+
  theme_minimal()+ggtitle("Scheduled versus Delayed Flights (Figure.1)")+
 xlab("day of month") + ylab("number")+ theme(plot.title = element_text(size=10))
```

Figure.2 Delayed versus Cancelled Flights

```r
# time series plot for December
# convert into long format
df_plot2 <- select(merged_date, DAY_OF_MONTH, prob1, prob2)
colnames(df_plot2)<-c("DAY_OF_MONTH","delayed","cancelled")
plot2<- melt(df_plot2,  id.vars = 'DAY_OF_MONTH', variable.name = 'status')
ggplot(plot2, aes(DAY_OF_MONTH, value)) +
  geom_line(aes(colour = status))+
  theme_minimal()+ggtitle("Delayed versus Cancelled Flights (Figure.2)")+
  xlab("day of month") + ylab("probability")+ theme(plot.title = element_text(size=10))
```

Correlation Coefficients

```r
# correlation total ~ delayed, total ~ canceled
cor(merged_date$freq,merged_date$prob1,method = "pearson", use = "complete.obs")
```

```
## [1] 0.5362778
```

```
# the correlation coefficient is 0.5352778
```

```
cor(merged_date$freq,merged_date$prob2,method = "pearson", use = "complete.obs")
```

```
## [1] 0.2340273
```

```
# 0.2340273
```

Figure.3 Delayed Flights by Major Carriers

```
# data visualization
# monthly flight by carriers, drop NA here since canceled flight is assigned to NA in df
# install.packages("wesanderson") for color palettes
library(wesanderson)

airline1 <- na.omit(airline)
ggplot(airline1,
       aes(x = OP_UNIQUE_CARRIER,fill = DELAY)) +
  geom_bar(position = "stack")+
  theme_minimal()+ggtitle("Delayed Flights by Major Carriers (Figure.3)")+
  xlab("major carriers") +
  ylab("number")+ scale_fill_manual(values=wes_palette(n=4, name="FantasticFox1"))+
  theme(plot.title = element_text(size=10))
```

Figure.4 FLight Delays of the Week

```
ggplot(airline1,aes(x = DAY_OF_WEEK, fill = DELAY)) +
  geom_bar(position = "stack")+
  theme_minimal()+ggtitle("FLight Delays of the Week (Figure.4)")+xlab("day of week") +
  ylab("number")+scale_fill_manual(values=wes_palette(n=4, name="Zissou1"))+
  theme(plot.title = element_text(size=10))
```

Model

```
# drop NA
flight <- na.omit(airline)

# set seed and randomly sample n=6000
set.seed(5849)
dff <- flight[sample(1:nrow(flight), 6000),]

# set categorical variables as factors
dff$DAY_OF_MONTH <- as.factor(dff$DAY_OF_MONTH)
dff$DAY_OF_WEEK <- as.factor(dff$DAY_OF_WEEK)
dff$OP_UNIQUE_CARRIER <- as.factor(dff$OP_UNIQUE_CARRIER)
dff$DEP_DEL15 <- as.factor(dff$DEP_DEL15)
dff$CANCELLED <- as.factor(dff$CANCELLED)
dff$ORIGIN_AIRPORT_ID <- as.factor(dff$ORIGIN_AIRPORT_ID)
```

```
library(brms)
set.seed(2020)
model <- brm(DEP_DEL15 ~ DAY_OF_WEEK +OP_UNIQUE_CARRIER+AIR_TIME,data=dff,seed=2020,
             family=bernoulli(),control=list(adapt_delta=0.90))
```

```
summary(model)
```

```
##  Family: bernoulli 
##   Links: mu = logit 
## Formula: DEP_DEL15 ~ DAY_OF_WEEK + OP_UNIQUE_CARRIER + AIR_TIME 
##    Data: dff (Number of observations: 6000) 
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
## 
## Population-Level Effects: 
##                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept              -1.60      0.19    -1.97    -1.23 1.00      679     1217
## DAY_OF_WEEK2           -0.24      0.11    -0.47    -0.02 1.00     2418     2913
## DAY_OF_WEEK3           -0.25      0.12    -0.48    -0.02 1.00     2292     2302
## DAY_OF_WEEK4           -0.49      0.12    -0.72    -0.25 1.00     2050     2591
## DAY_OF_WEEK5           -0.21      0.11    -0.43     0.01 1.00     2202     2958
## DAY_OF_WEEK6           -0.21      0.12    -0.45     0.03 1.00     2269     2533
## DAY_OF_WEEK7           -0.05      0.11    -0.27     0.16 1.00     2274     2622
## OP_UNIQUE_CARRIERAA     0.10      0.20    -0.30     0.49 1.00      669     1303
## OP_UNIQUE_CARRIERAS     0.34      0.24    -0.14     0.82 1.00      864     1530
## OP_UNIQUE_CARRIERB6     0.95      0.22     0.52     1.40 1.00      814     1360
## OP_UNIQUE_CARRIERDL    -0.09      0.21    -0.49     0.33 1.00      728     1241
## OP_UNIQUE_CARRIEREV     0.50      0.30    -0.06     1.10 1.00     1352     2071
## OP_UNIQUE_CARRIERF9     0.72      0.28     0.16     1.28 1.00     1221     2111
## OP_UNIQUE_CARRIERG4     0.72      0.28     0.18     1.25 1.00     1134     1877
## OP_UNIQUE_CARRIERHA    -0.46      0.50    -1.51     0.44 1.00     2162     2272
## OP_UNIQUE_CARRIERMQ     0.32      0.24    -0.13     0.77 1.00      862     1535
## OP_UNIQUE_CARRIERNK    -0.08      0.30    -0.69     0.48 1.00     1175     1814
## OP_UNIQUE_CARRIEROH     0.34      0.24    -0.13     0.83 1.00      905     1646
## OP_UNIQUE_CARRIEROO     0.15      0.20    -0.24     0.56 1.00      748     1401
## OP_UNIQUE_CARRIERUA     0.23      0.21    -0.17     0.65 1.00      743     1522
## OP_UNIQUE_CARRIERWN     0.64      0.19     0.27     1.03 1.00      656     1155
## OP_UNIQUE_CARRIERYV     0.27      0.26    -0.26     0.78 1.00     1107     1920
## OP_UNIQUE_CARRIERYX     0.19      0.24    -0.27     0.64 1.00      906     1740
## AIR_TIME                0.00      0.00    -0.00     0.00 1.00     4203     3069
## 
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Model Diagonstics

Figure.5: Posterior Predictive Graph

```
pp_check(model,nsamples = 100)
```

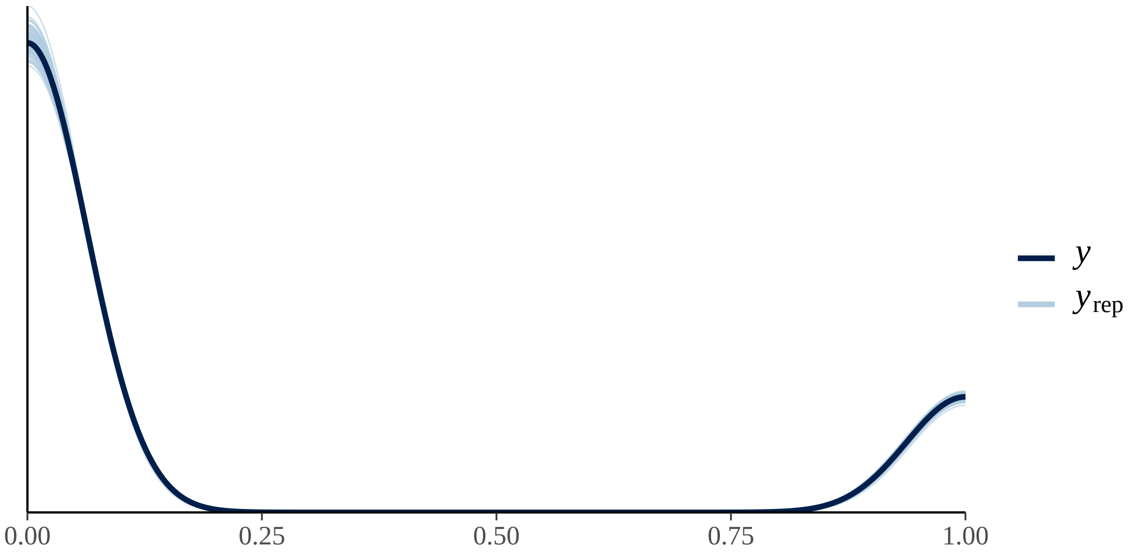
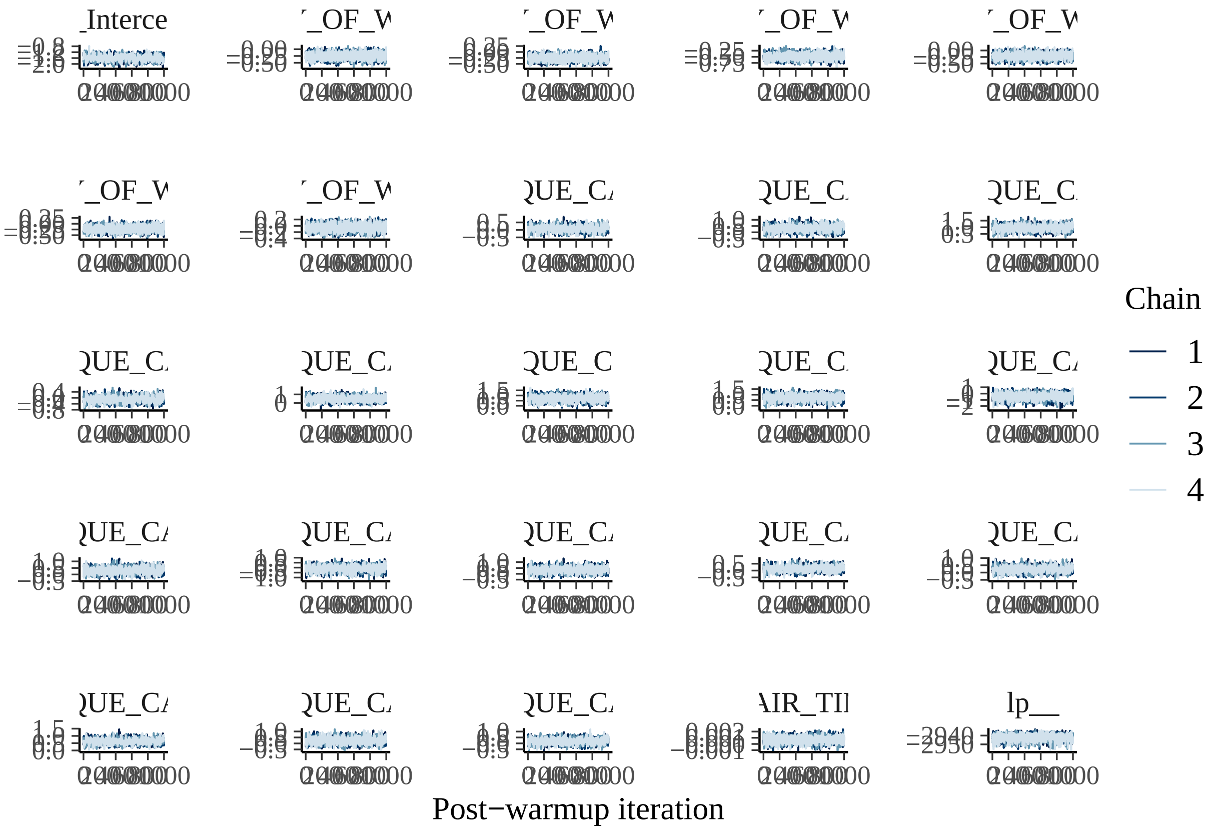


Figure.6: Non-convergence Plot

```r
library(bayesplot)
mcmc_trace(model)+xlab("Post-warmup iteration")
```



Shinystan

```r
# library(shinystan)
# launch_shinystan(model), the results will be shown on the website
```

Model Results

Table.1 Population-Level Effects

```r
# make a table for the output
library(kableExtra)
yz_table <- data.frame("estimated coefficients" =
                         c("-1.60","-0.24","-0.25","-0.49","-0.21",
                           "-0.21","-0.05", "0.10","0.34", "0.95","-0.09",
                           "0.50","0.72","0.72","-0.46","0.32","-0.08","0.34",
               "0.15","0.23","0.64","0.27","0.19","0.00"),
    "lower 95%CI"=c(" -1.97","-0.47","-0.48","-0.72","-0.43","-0.45","-0.27",
                    "-0.30","-0.14","0.52","-0.49","-0.06","0.16","0.18","-1.51",
                    "-0.13","-0.69","-0.13","-0.24","-0.17","0.27","-0.26","-0.27","0.00"),
    "upper 95%CI"=c("-1.23","-0.02","-0.02","-0.25","0.01","0.03","0.16","0.49","0.82",
                    "1.40","0.33","1.10","1.28","1.25","0.44","0.77","0.48","0.83",
                    "0.56","0.65","1.03","0.78","0.64","0.00"))
row.names(yz_table) <- c("Intercept","Tuesday","Wednesday","Thursday","Friday",
                         "Saturday","Sunday","American Airlines (AA)","Alaska Airlines (AS)",
                         " JetBlue Airways (B6)","Delta Air Lines (DL)",
                         "Atlantic Southeast Airlines (EV)","Frontier Airlines (F9)",
                         "Allegiant Air (G4)","Hawaiian Airlines (HA)","Envoy Air (MQ)",
                         "Spirit Airlines (NK)","Comair (OH)","Sky West Airlines (OO)",
                         "United Airlines (UA)","Southwest Airlines (WN)",
                         "Mesa Airlines (YV)","Repblic Airways (YX)","AirTime")

table <- yz_table %>%
  kbl(caption = "Population-Level Effects") %>%
  kable_styling(bootstrap_options = "striped", position = "center",full_width = F) %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```