Stat 421 Final Project: Part 2
David Kim & Ji Hoon Woo
December 20, 2018

"Why My Flight Has Been Delayed Again"

1. Overview of Research Questions & Results

1.1 Do the three U.S. largest airlines: American, Delta & United have a lower probability of departure delay than the low-cost airlines due to their high-ticket price?

- Regardless of the size of airline or the price of ticket, the odds of having departure delay vary by airline for different reason.

1.2 Which time of the day has the largest departure delay?

- As time goes by, we are more likely to observe the departure delay and evening is the peak time of moment.

1.3 How much does the airtime affect the departure delay of American Airline for 1-hour increase?

- Interestingly, there is a negative relationship with the time of increase in airtime such that you are less likely to have the departure delay with flights of long duration.

2. Motivation & Background

Our group is originally from South Korea and we travel often by airplane for both domestic and international flights. Last month, we had a chance to visit Los Angeles for job interview and there was again a departure delay on the way back to Madison from Saint Paul Airport in Minneapolis, Minnesota. Even though we were tired after the job interview, we could not help waiting for the plane to be ready. Our complaint eventually intrigued our curiosity of why departure delay occurs every day.

According to Peterson et al, the author of "The Economic Cost of Airline Flight Delay" (2013), estimated total hours of delay of passenger airplane is about 468.2 million hours in 2007. In fact, this is huge amount of time and considering the flight delay cost per minute is $61, we are making huge economic loss. In addition, from 2002 to 2007, while the number of total airplanes that flew in the United States were increased by 40.7 percent, the number of airplanes that arrived 15 minutes than the scheduled itinerary time were increased by 106.4 percent. This is a strong evidence supporting our statement that although the number of flights is increasing fast, late arrival of flight is more probable than expected. Then we began to investigate factors that could possibly affect the departure delay and question ourselves ways to reduce the delay time.

3. Dataset

The dataset we will be using contains data on Flight information and is available at https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time. In order to conduct an accurate analysis, it is important to use the verified data that only contains non-estimated values. Therefore, we were able to find the dataset from United States Department of Transportation for our project. Since there are more than 87,000 airplanes depart from the United States per day, we narrow it down to only six major airlines flights that depart in January. The dataset contains 9 variables and 33,208 rows.

We chose airline, state of departure, departure time, airtime, distance to destinations, existence of malfunctioning aircraft, weather indicator, and late aircraft to be our 8 explanatory variables to explain the response variable which is departure delay time that is categorical variable. We only chose 6 main carriers for our airline variable because these are the only airlines that has consistent number of flights compared to last year. Departure hour is used to see whether the peak hour or sunlight has any impact on causing delay. In order to see the impact of number of passengers in each flight, we add distance and airtime variable, as most of long-distance flight tend to be operated with airplane that has high capacity which causes longer period to board passengers. We also included 3 indicator variables: weather, carrier and late aircraft. Weather variable is included to check the impact of bad weather such as rain or snow which might slow down the operation at the airport. Late aircraft variable is the last one that we consider because there are many cases when departure delay is caused due to the late arrival of aircraft from the previous flight.

4. Methodology (Algorithm & Analysis)

4.1 Data Cleansing

The data contains 156 missing values in variable "Airtime" and we transformed them into "0" before an analysis begins. Also, we modified the data to the class "factor" or "integer" so that we may be ready to perform the logistic regression with the variable "Depdelay" as our dependent variable.

4.2 Model Selection

We tried to find the best model to obtain the right balance between parsimony and goodness of fit. By running and comparing both methods: Akaike's Information Criterion (AIC) & Bayesian Information Criterion (BIC), our team concluded that 1st AIC model is most ideal with 8 exploratory variables and its value of 28938.26.

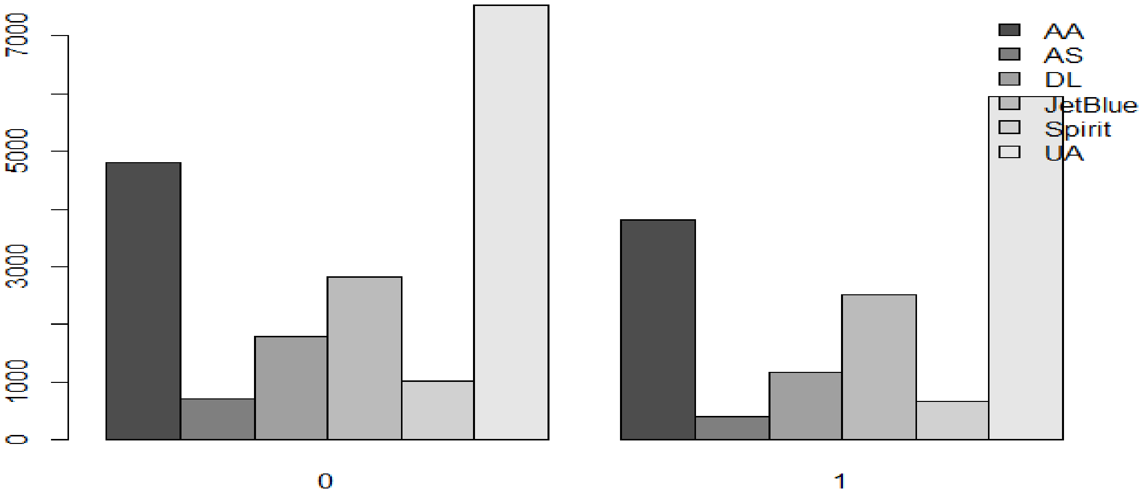
4.3 Statistical Analysis

- Based on our model, we calculated the odds of 6 different airlines to investigate the frequency of departure delay. Then we computed the odd ratio of 2 certain airlines and its confidence interval to guarantee the range of our values.

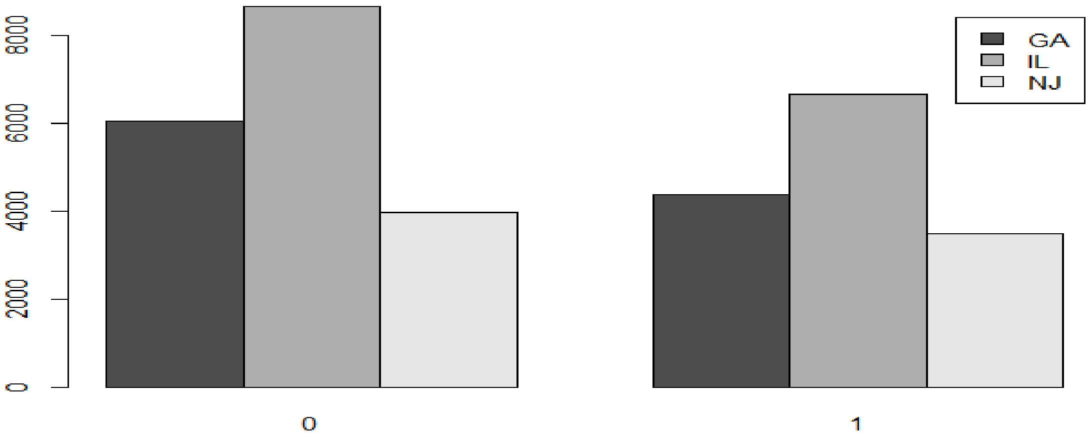
- In addition, we investigated another odd in terms of departure time. Our departure time is divided into 4-time intervals: Early Morning, Morning, Afternoon & Evening. Then we again calculated the ratio of each time of the day and defined the confidence interval.

- Next, we compared the odds based on the airport. We collected data from 3 largest airports that have the most departure delays (O'Hare, Newark & Hartsfield-Jackson) in order to find out which airport have higher odds of departure delay than the others.
- For our last investigation, we ran Poisson Regression to identify the coefficients of each variable. Using this data, we calculate the percent change of departure delay for airtime increase.

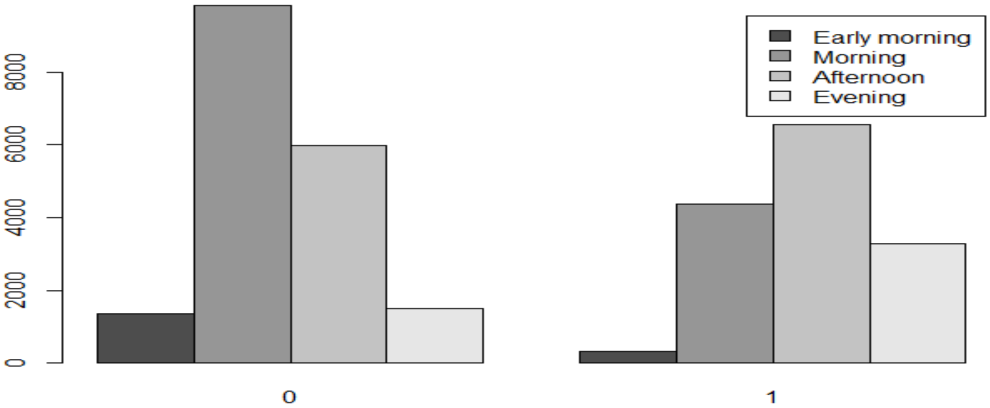
5. Results



This barplot indicates the number of departure delay based on the airline carrier. United Airlines has the largest number of aircrafts operation on a daily basis. We normally presume that large airlines are more likely to have less departure delay due to their higher price ticket, however the result was different than expected. Jet Blue Airlines had the highest odds that there were 89 delay flights for every 100. This is followed by American Airlines (AA) with 80 delay flights out of 100, United Airlines with 79 out of 100. Interestingly, one of the largest airline carriers, DELTA (DL), has quite low odds with 65 out of 100 followed by Alaska Airlines (AS) with the ratio 57 out of 100. When we calculated the odds ratio between JetBlue and Alaska Airlines, the odds are 1.568905 times the estimated odds for Alaska Airlines within 95% confidence interval between 1.371696 and 1.794326. In other words, "**for every 100 delayed flights for Alaska Airlines, we would expect 157 delayed flights for JetBlue**."



This is another plot describing the number of departure delay based on 3 largest airports. Before an analysis, our team assumed that there will not be any large difference among 3 airports because these 3 airports are well-known for having frequent delays. Each acronym represents the state where the airports are located in. Hartsfield-Jackson in Georgia, Atlanta, O'Hare in Chicago, Illinois and Newark in New Jersey. Among the airports, Newark has largest odds that there are 88 delay flights for every 100 flights. O'Hare had 77 out of 100 and Hartsfield-Jackson with 72 out of 100. When comparing the odds ratio between Newark and O'Hare, we can interpret it as "**For every 100 delayed flights at O'Hare, we would expect to see 115 at Hartsfield-Jackson within 95% confidence interval between 104 and 126**."

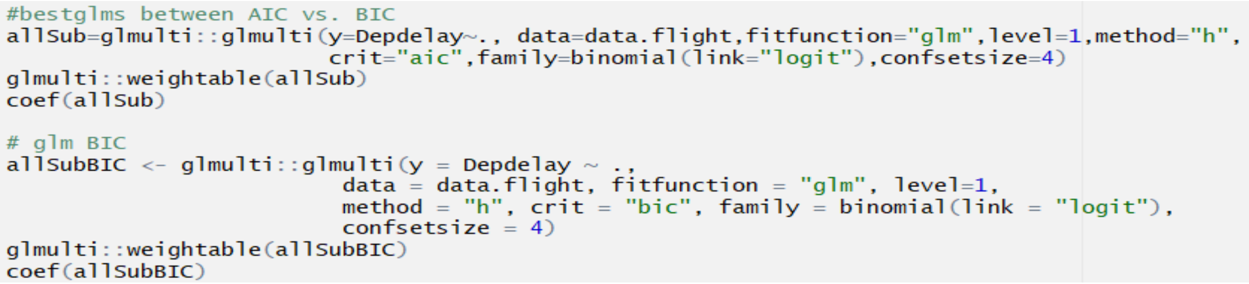


Our last barplot represents the number of departure delay based on the time of the day. There are only few flights flying in the early morning and our result also indicates that the odds of having flight delay is only .24 meaning that there are only 24 flights out of 100. As the day goes by, however, the odds exponentially increase that the odds are .44 in morning, 1.1 in the afternoon and eventually 2.18 in the evening. This means that there are 218 delay flights for every 100 flights. When we compare the odds with the odds in the early morning, it is 8.93 times the odds in the early morning such that "**we would expect 893 delayed flights in the evening for every 100 delayed flights in the morning within 95% confidence interval between 779 and 1,020**".

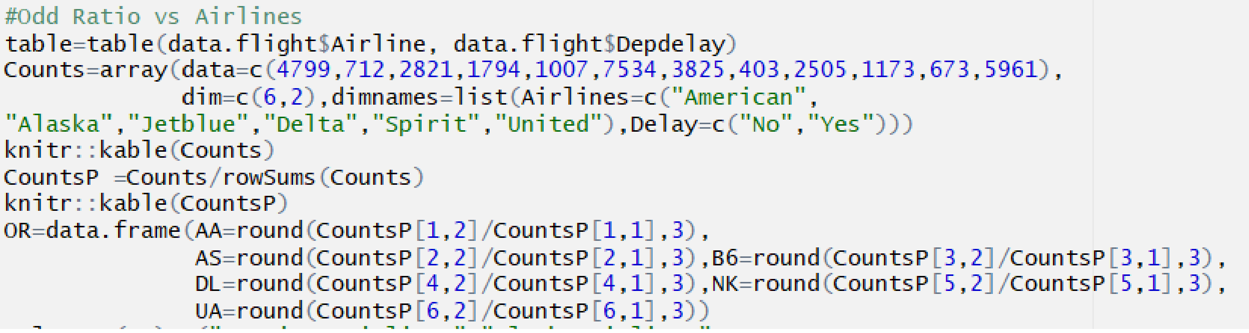
For our final result, our team calculated the departure delay of American Airline for airtime increase. As the airtime increases, there is an interesting fact that departure delay decreases by 14.58% from 3 hours of airtime to 4 hours. Thus, there is negative correlation between airtime and departure delay that as the airtime increases, the percentage of departure delay decreases.

Based on our result, we found an interesting tip or fact that might benefit frequent airline users. Having a flight in O'Hare with Alaska Airlines might be the greatest option for those who really hate flight delays or would like to avoid any potential flight departure by chance.

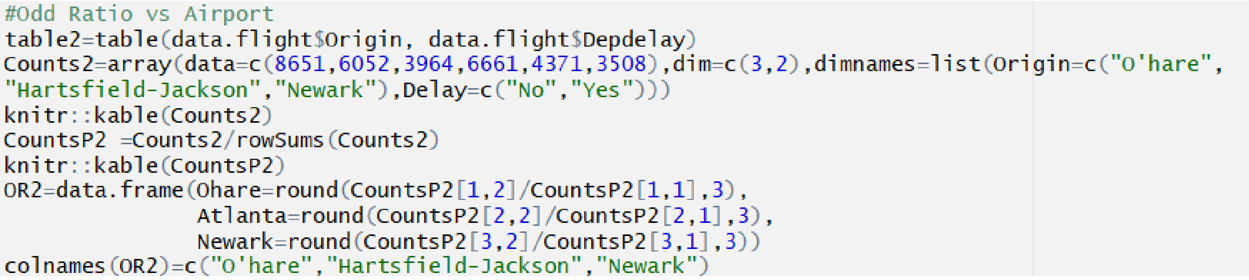
6. Reproducing Your Results



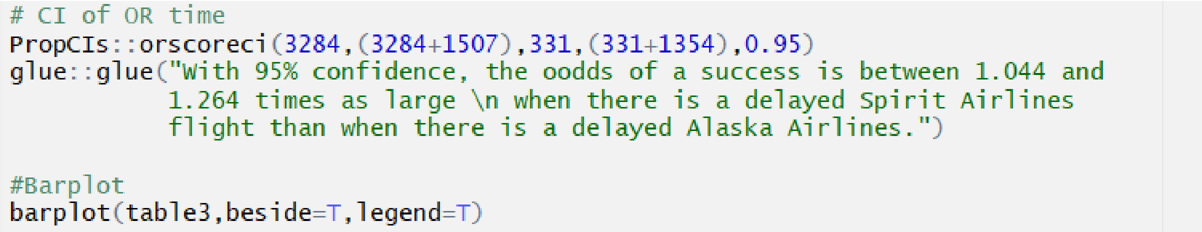
```
#bestglms between AIC vs. BIC
allSub=glmulti::glmulti(y=Depdelay~., data=data.flight,fitfunction="glm",level=1,method="h",
                        crit="aic",family=binomial(link="logit"),confsetsize=4)
glmulti::weightable(allSub)
coef(allSub)

# glm BIC
allSubBIC <- glmulti::glmulti(y = Depdelay ~ .,
                        data = data.flight, fitfunction = "glm", level=1,
                         method = "h", crit = "bic", family = binomial(link = "logit"),
                         confsetsize = 4)
glmulti::weightable(allSubBIC)
coef(allSubBIC)
```



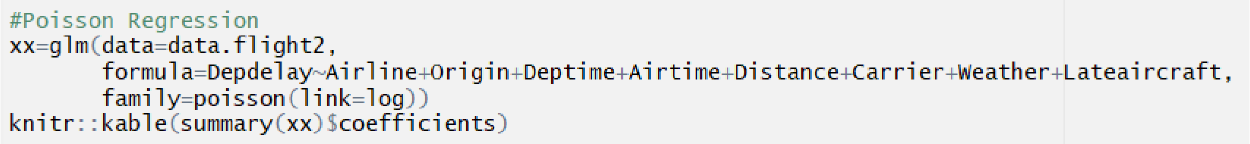
```
#Odd Ratio vs Airlines
table=table(data.flight$Airline, data.flight$Depdelay)
Counts=array(data=c(4799,712,2821,1794,1007,7534,3825,403,2505,1173,673,5961),
             dim=c(6,2),dimnames=list(Airlines=c("American",
"Alaska","Jetblue","Delta","Spirit","United"),Delay=c("No","Yes")))
knitr::kable(Counts)
CountsP =Counts/rowSums(Counts)
knitr::kable(CountsP)
OR=data.frame(AA=round(CountsP[1,2]/CountsP[1,1],3),
               AS=round(CountsP[2,2]/CountsP[2,1],3),B6=round(CountsP[3,2]/CountsP[3,1],3),
               DL=round(CountsP[4,2]/CountsP[4,1],3),NK=round(CountsP[5,2]/CountsP[5,1],3),
               UA=round(CountsP[6,2]/CountsP[6,1],3))
```



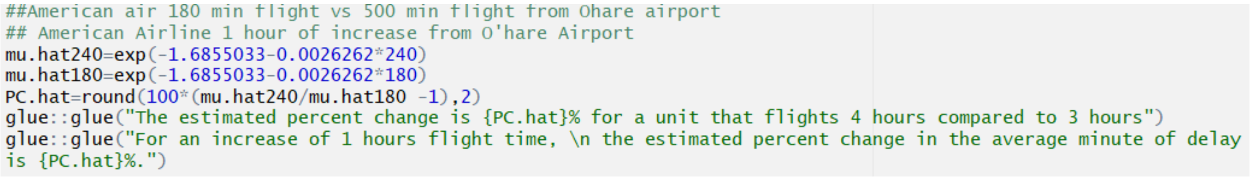
```
#Odd Ratio vs Airport
table2=table(data.flight$Origin, data.flight$Depdelay)
Counts2=array(data=c(8651,6052,3964,6661,4371,3508),dim=c(3,2),dimnames=list(Origin=c("O'hare",
"Hartsfield-Jackson","Newark"),Delay=c("No","Yes")))
knitr::kable(Counts2)
CountsP2 =Counts2/rowSums(Counts2)
knitr::kable(CountsP2)
OR2=data.frame(Ohare=round(CountsP2[1,2]/CountsP2[1,1],3),
               Atlanta=round(CountsP2[2,2]/CountsP2[2,1],3),
               Newark=round(CountsP2[3,2]/CountsP2[3,1],3))
colnames(OR2)=c("O'hare","Hartsfield-Jackson","Newark")
```



```
# CI of OR time
PropCIs::orscoreci(3284,(3284+1507),331,(331+1354),0.95)
glue::glue("With 95% confidence, the oodds of a success is between 1.044 and
          1.264 times as large \n when there is a delayed Spirit Airlines
          flight than when there is a delayed Alaska Airlines.")

#Barplot
barplot(table3,beside=T,legend=T)
```



```
#Poisson Regression
xx=glm(data=data.flight2,
       formula=Depdelay~Airline+Origin+Deptime+Airtime+Distance+Carrier+Weather+Lateaircraft,
       family=poisson(link=log))
knitr::kable(summary(xx)$coefficients)
```



```
##American air 180 min flight vs 500 min flight from Ohare airport
## American Airline 1 hour of increase from O'hare Airport
mu.hat240=exp(-1.6855033-0.0026262*240)
mu.hat180=exp(-1.6855033-0.0026262*180)
PC.hat=round(100*(mu.hat240/mu.hat180 -1),2)
glue::glue("The estimated percent change is {PC.hat}% for a unit that flights 4 hours compared to 3 hours")
glue::glue("For an increase of 1 hours flight time, \n the estimated percent change in the average minute of delay
is {PC.hat}%.")
```

7. Collaboration & Reflection

Before the project begins, our team did not come up with any intuitive questions of why the departure delay occurs in the airport. We just followed by the announcement that the airplane has been delayed due to whatever reason. Throughout this project, out team learned how to think critically and statistically based on the causality of the event. We also had an opportunity to implement our statistical analysis learnt from this course to the real-world data. It was very interesting that our analysis implies the results which might benefit others.

Overall, we did not have any outside assistance but the advice from Instructor Michael Holt and 2 other peer reviews. Our team members spent approximately 7 hours individually for Part 1 and another 15 hours for Part 2 so that we could increase the quality of our experiment. One reflection that our team made from this project is that if there was a chance for us to present our overall research, it would be much productive and instructive for all the students taking this class.

8. Reference

Peterson, Everett B. “The Economic Cost of Airline Flight Delay.” *The Economic Cost of Airline Flight Delay*, Jan. 2013, doi:10.1108/02640470410520203.