Flight Delay Analysis Report

ANLY 511 - Final Project 2019

December 9, 2019

Group 2: Xueyan Liu, Yiran Liu, Anderson Monken, and Shiying Sun

# Introduction

What’s up with flight delays? Being delayed is one of the most frustrating aspects of flying. You get almost no information, and you are helpless to the will of the airline carrier. Sometimes it is due to weather, faulty equipment, or even software failures. We have all experienced significant delays during flying that have caused severe inconveniences. In the first half of 2018 almost 4.5 million passengers had a flight cancelled, delayed, or overbooked[[1]](#footnote-1).

There were one-billion passengers on domestic and international flights in the United States in 2018[[2]](#footnote-2), and in October of 2019 that number came to 78.1 million[[3]](#footnote-3). This means that the return on investment to improving the efficiency and reliability of airline flights is substantial. Idle planes, overworked airline employees, and occupied airline gates are all causes of financial loss to airline companies, which run on thin margins to make profits. When the average minutely cost to airlines for a single plane is estimated to be $74.20[[4]](#footnote-4), even minor delays lead to significant lost revenue. Analysis by the FAA/Nextor estimated direct cost of delays to airlines and passengers was 28 billion dollars in 2018[[5]](#footnote-5).

From the perspective of the passenger, when one purchases a flight, they expect to be delivered to their destination safely and on time. While there is financial compensation for passengers who are delayed significantly by flights, missed conferences, being late to meetings, and not being at family events are all indirect costs that passengers bear when a flight gets delayed. Therefore, there is a strong consumer interest in knowing the likelihood of being on a delayed flight, and incorporating that probability into their calculus for the most efficient mode of transportation. To that end, this paper intends to predict scenarios that are most likely to lead to a delayed flight in New York City.

There is a strong history of researching the patterns in flight delays. According to previous research, flight departure delay is a very complex problem with substantial direct causal factors and many concealed indirect causal factors[[6]](#footnote-6). The indirect causal factors may include previous delayed flights, airports management and governmental issues. Thus our models does not have a perfect performance using direct variables.

# Analysis/Statistical Methods

## Data Source

## Processing

Data came from five distinct datasets containing information on airports, airplanes, airline carriers, flights, and weather. These datasets were merged together using common columns to build a master dataset.

The airplanes dataset was ultimately not used due to lack of data. While the year of the plane might serve as a proxy for its suitability for flying since newer planes likely would require fewer mechanical issues, due to excessive missing values the airplane variables were dropped from the final dataset.

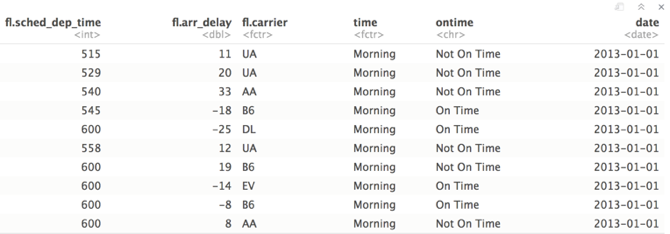
The flights dataset had a number of missing values for arrival or departure delays. These missing values were actually flights that were cancelled entirely. These values were recoded to a large delay of 5 hours in the dataset since we would expect those passengers to get onto a different flight eventually.

Variable binning was performed on a number of features in the dataset including time of day, arrival delay, departure delay, and more. For the chi-squared analysis on airline carriers, cut points were -90,0,5,30,60,120, Inf. The value counts for departure delays are shown by airline carrier.



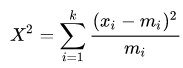
For probit, logit, and naive bayes analysis the departure delays and arrival delays were converted to boolean values for if a delay occurred or not, with 0 inclusive of the false group.

For Chi-squared test for Independence on time periods of a day and on-time rate of flights, the scheduled departure time is divided into four time periods, morning, afternoon, evening, and late night. The time periods range from 5am to 11am, 11am to 5pm, 5pm to 11pm, and 11pm to 5am. The new column named “time” is created for the time period values. The arrival delay values are used to determine if the flight is on time or not. The arrival delay value which is equal to or is smaller than 0, is classified as “On Time”. The arrival value which is smaller than 0 or is NA, which is cancelled flight, is classified as “Not On Time”. The new column named “ontime” is created for flight status. The new column “date” is combined and formatted from “year”, “month”, and “day” columns in flight dataset.



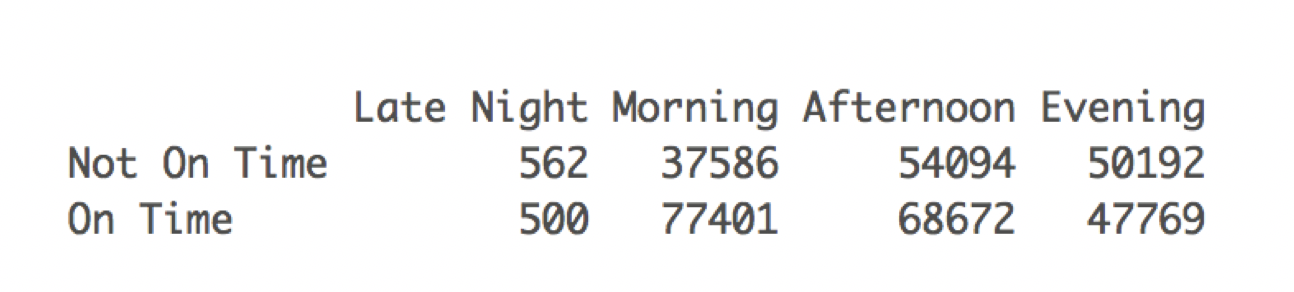
## Statistical Tests

Chi-squared tests for independence were performed to test the statistical significance of multiple relationships in the dataset. The chi-squared test for independence compares the relationship between two variables. If there is a relationship between them then knowing one will affect the prediction on the other. In this paper we separately tested if various independent variables such as airline carrier or time of day would affect the distribution of departure/arrival delays. The test statistic of interest is the chi-squared statistic, which is a square of differences of actual counts for each delay category / independent variable category combo.



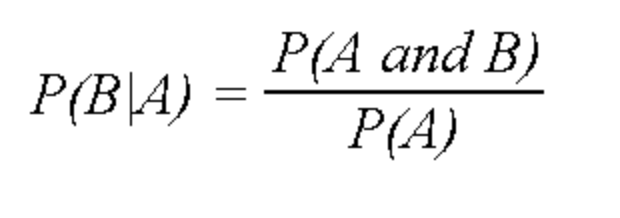
The null hypothesis for this statistical test is that there is no relationship between the independent variable and the outcome variable (delays). The alternative hypothesis is that there is a relationship between the two. All tests were taken at the 95% confidence level.

For Chi-squared test for Independence on time periods of a day and on-time rate of flights, the columns “time” and “ontime” created in data processing are used to perform the hypothesis test. The null hypothesis is that on-time rate and daytime periods are independent. On-time rate do not vary by daytime periods. The alternative hypothesis is that on-time rate and daytime periods are dependent. On-time rate do vary by daytime periods. Below is the table containing counts of flight “not on time” times and counts of flight “on time” times for each of time periods.



Conditional Probability

Conditional probability is the probability of an event occurs, given that other events have already occurred. The probability of event B given event A is showing as below, which is equal to the probability of both event A and B occurs divided by the probability of event A.

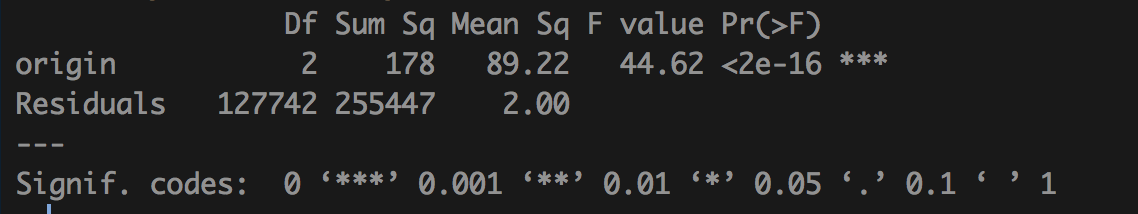


The conditional probability will be used to analyze the on-time rate for each time period during a day. Additionally, each airline’s on-time perform within each time periods will be analyzed by using conditional probability as well. The columns scheduled departure time, arrival delay, carrier and the date will be used for this conditional probability analysis.

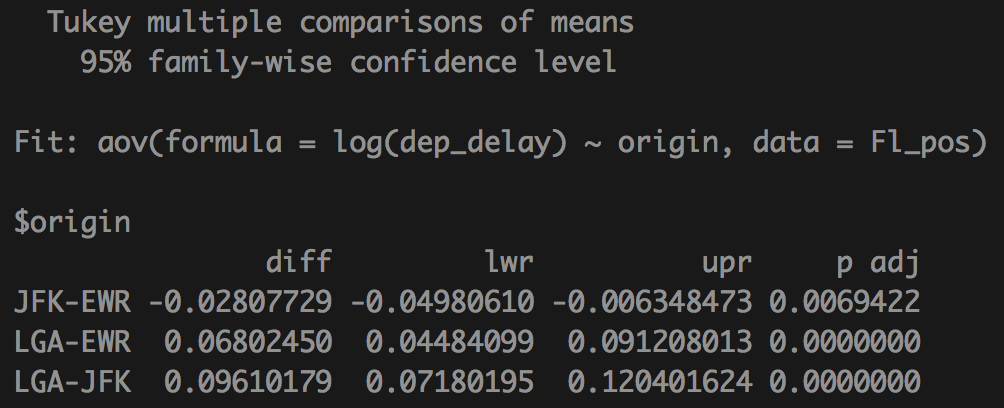
ANOVA

Since the departure delay variable is severely skewed, our group transformed it into log(departure\_delay) before applying ANOVA tests.

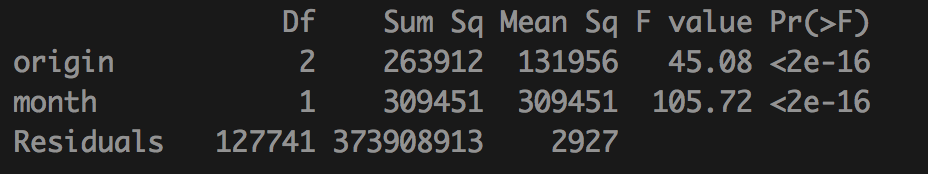
In order to determine whether the difference on departure delay is statistically significant enough or not among three different origin airports. One way ANOVA test was performed on the departure delay variable and the origin airport variable. The result is as follows: With P value significantly smaller than 0.1%, it is appropriate to conclude that there is a significant difference on departure delay among origin airports.



Then a pairwise ANOVA test was applied. The result is as follows: The confidence interval of each pair indicates our test have 95% chance to say that the true population difference between each origin airport on departure delays will fall in the interval between lwr and upr.



In order to determine whether the difference on departure delay is statistically significant enough or not among three different origin airports and month. Another two way ANOVA test was performed on the departure delay variable, the origin airport variable and the month variable. The result is as follows: With both P value significantly smaller than 0.1%, it is appropriate to conclude that there is a significant difference on departure delay among origin airports and different month.



## Multiple Linear Regression

In order to determine the magnitude of the effect of the independent variables in our study, multiple linear regression (MLR) analysis was performed on the departure and arrival delay dependent variables. The formula for the regression takes the form of a constant plus multiple independent variables with a weight and an error term.

The optimizer for the multiple linear regressions conducted is the least-squares approximation, which minimizes the squared difference between the underlying data and the linear approximation of the data. The error terms for all observations should be minimized and evenly distributed across the fitted values. Another important measure is the R^2 measure, which explains the amount of variation in the data that is explained by the regression. Each individual variable is assigned a p-value for its significance.

Multiple transformations were performed on the data for MLR to better normalize the data and get stronger prediction power from the regression. The first transformation was a modified log transformation of the form f(x) = log(x + 1 - min(x)). This modification was made because negative values were present in the data and the log function diverges to negative infinity as x approaches 0. The second transformation made was a modified inverse function f(x) = 1/(1+5\*x). This modification was made because some observations were 0, where 1/x is undefined. There were also issues with the regression returning errors that led to the multiplication of x.

Success of the MLR is determined by R^2 value, which serves as the predictive power for the delay expected for a flight that meets the characteristics of a given flight.

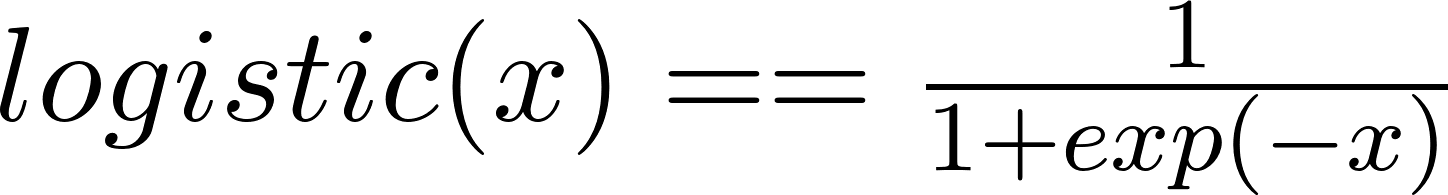
## Probit Model

The binary delays were used as response variables for the probit model regression. This model estimates the probability of a flight delay occurring. The estimator for this regression is the maximum likelihood procedure. It is considered part of the generalized linear model family. The prediction from the model are values [0,1] representing the probability of delay. These values were then converted to binary 1 or 0 if the model was >50% probability of delay or <=50% probability. The model performance is judged by Akaike's information criterion (AIC) and model accuracy. The data is split into 80% training and 20% test data, and model accuracy is determined by the correct predictions from the model on the testing data.

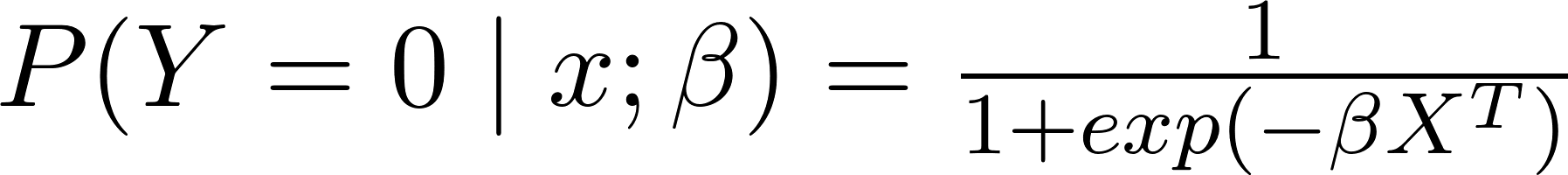
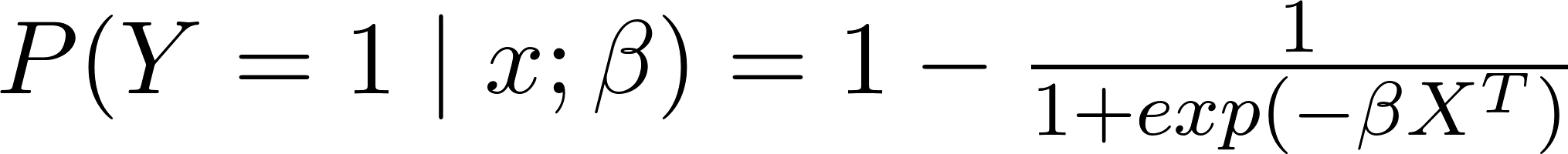
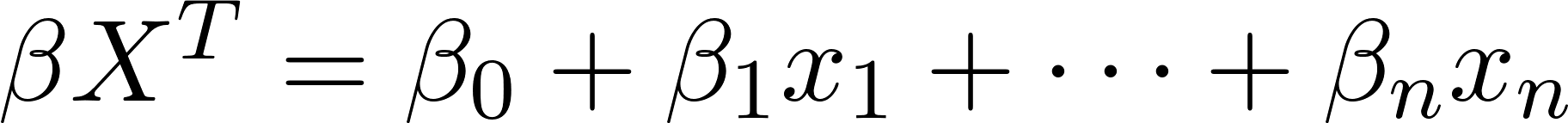
## Logistic Regression Model

We also implement the logistic regression, a classification model, to finding the probabilities of flight delay. Logistic regression is a widely used machine learning model[[7]](#footnote-7) because it is very efficient. It doesn’t need normalized the inputs, is cheap to implement, and is easy to interpret and validate the results. However, logistic regression will also be impacted by the linear dependent variables and noisy data.

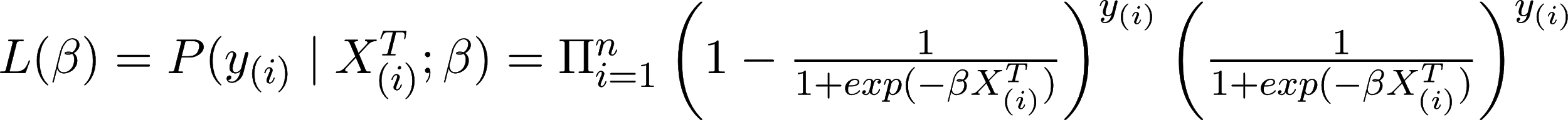
Compared to linear regression which predict a continuous outcome, logistic regression predicts the classification and returns a discrete outcome (‘1’ or ‘0’) of a binary question, such as “will tomorrow be raining” or “whether the flight will be delayed?”. And instead of fitting a straight line [[8]](#footnote-8), logistic regression try to fit the logistic function, where the logistic function is defined as:

[](https://www.codecogs.com/eqnedit.php?latex=logistic(x)%20%3D%20%3D%5Cfrac%7B1%7D%7B1%2Bexp(-x)%7D%250).

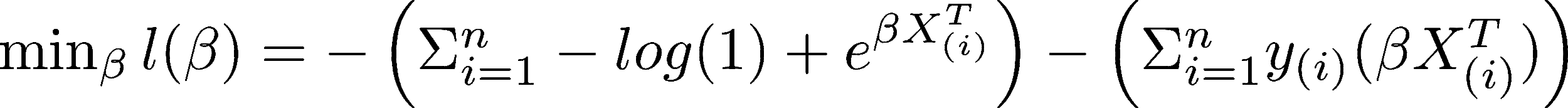
In mathematically, logistic regression estimates the parameter of the conditional expectations of y:

[](https://www.codecogs.com/eqnedit.php?latex=P(Y%20%3D%200%20%5Cmid%20x%20%3B%20%5Cbeta%20)%20%3D%20%5Cfrac%7B1%7D%7B1%2Bexp(-%20%5Cbeta%20X%5ET)%7D%250) and [](https://www.codecogs.com/eqnedit.php?latex=P(Y%20%3D%201%20%5Cmid%20x%20%3B%20%5Cbeta%20)%20%3D%201-%5Cfrac%7B1%7D%7B1%2Bexp(-%20%5Cbeta%20X%5ET)%7D%250) where [](https://www.codecogs.com/eqnedit.php?latex=%5Cbeta%20X%5ET%20%3D%20%5Cbeta_0%2B%5Cbeta_1x_1%2B%5Cdots%20%2B%20%5Cbeta_nx_n%250),

and allows us to predict y with given X. In order to figure out the model parameters, maximum likelihood estimation method is introduced. Thus, the likelihood function is

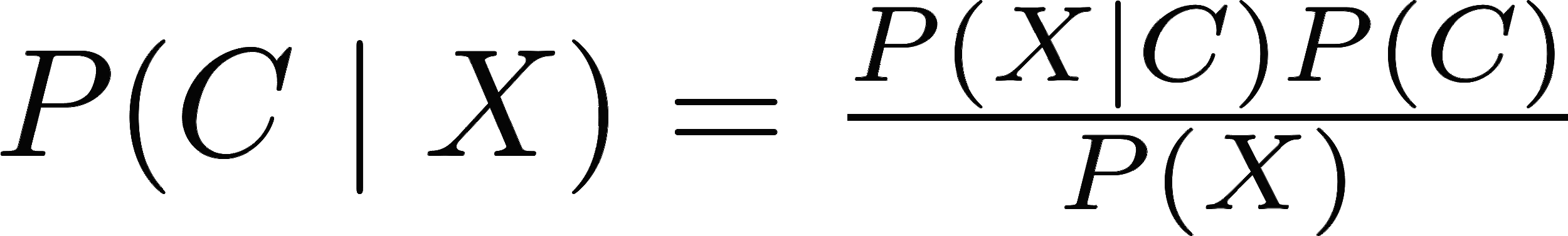
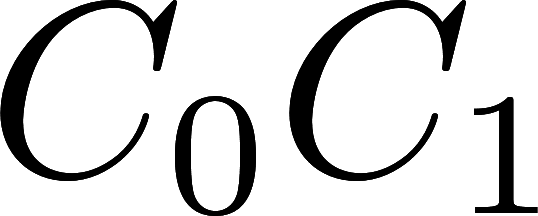
[](https://www.codecogs.com/eqnedit.php?latex=L(%5Cbeta)%20%3D%20P(y_%7B(i)%7D%20%5Cmid%20X_%7B(i)%7D%5ET%3B%5Cbeta%20)%20%3D%20%5CPi_%7Bi%3D1%7D%5En%20%5Cleft%20(1-%5Cfrac%7B1%7D%7B1%2Bexp(-%5Cbeta%20X_%7B(i)%7D%5ET)%7D%20%5Cright)%5E%20%7By_%7B(i)%7D%7D%20%5Cleft(%5Cfrac%7B1%7D%7B1%2Bexp(-%5Cbeta%20X_%7B(i)%7D%5ET)%7D%20%5Cright)%5E%7By_%7B(i)%7D%7D%250) [[9]](#footnote-9)

which represents how possible the model parameters fit the given data have given. Then we maximize the likelihood function to find the parameters by Newton's method. Overall the logistic regression is trying to solve the program:[[10]](#footnote-10)

[](http://www.texrendr.com/?eqn=%5Cmin_%5Cbeta%20%20l(%5Cbeta)%20%3D%20%20-%20%5Cleft(%7B%5CSigma_%7Bi%3D1%7D%5En%7D%20-log(1)%2Be%5E%7B%5Cbeta%20X_%7B(i)%7D%5ET%7D%20%20%5Cright)-%5Cleft(%5CSigma_%7Bi%3D1%7D%5En%20y_%7B(i)%7D(%5Cbeta%20X_%7B(i)%7D%5ET)%20%5Cright)%250)

Because logistic regression relies on the Maximum likelihood estimates rather than minimize the variance, we can not apply R squared to measure the goodness of fit of logistic regression model. Instead, we will use the Pseudo R-Squared, also known as ‘McFadden's R squared’[[11]](#footnote-11), which define as [](https://www.codecogs.com/eqnedit.php?latex=R_%7BMcFadden%7D%5E2%20%3D%201-%20%5Cleft(%5Cfrac%20%7Blog(%5Ctext%7Bthe%20maximized%20likelihood%20value%20from%20the%20current%20fitted%20model%7D)%7D%7Blog(%5Ctext%7Bthe%20corresponding%20value%20but%20for%20the%20null%20model%7D)%7D%20%5Cright)%250). The pseudo R squared suggests the “level of improvement over the intercept model offered by the full model”[[12]](#footnote-12) ; the higher pseudo R squared indicate the better model. [[13]](#footnote-13)

## Naive Bayes

Naive Bayes is another classification model, same as logistics regression, it returns a binary discrete outcome. Naive Bayes method is built on the Bayes’ Theorem which incorporates the conditional probability(1) and the formula is defined as [](http://www.texrendr.com/?eqn=P(C%20%5Cmid%20X)%20%3D%20%5Cfrac%20%7BP(X%20%5Cmid%20C)P(C)%7D%7BP(X)%7D%250)[[14]](#footnote-14).where X, representing the [](http://www.texrendr.com/?eqn=x_1%2Cx_2%2C%20%5Cdots%2C%20x_n%250), denotes different variables that will impact the C; and C representing[](http://www.texrendr.com/?eqn=C_0%20%26%20C_1%250), denotes two possible outcomes of a certain event. In this case we use the Naive Bayes method to compute the probability of flight delays by the function:

P(delay \text{ & } conditions) = P(condition_1 \mid delay) \cdots  P(condition_n \mid delay)P(delay)

The advantages of this method are obvious: it is simple and fast, and the results are intuitive. But the disadvantages are also obvious. It assumes that all variables are independent, which is extremely hard to reach; and all the variables should be categorical variables.[[15]](#footnote-15) Thus in order to implement the Naive Bayes method, we need to transfer continuous variables to the categorical variables, bining all the continuous variables into four quartiles. And we marked delay flight as ‘1’ and on time flight as ‘0’ as well, just as we did in previous methods.

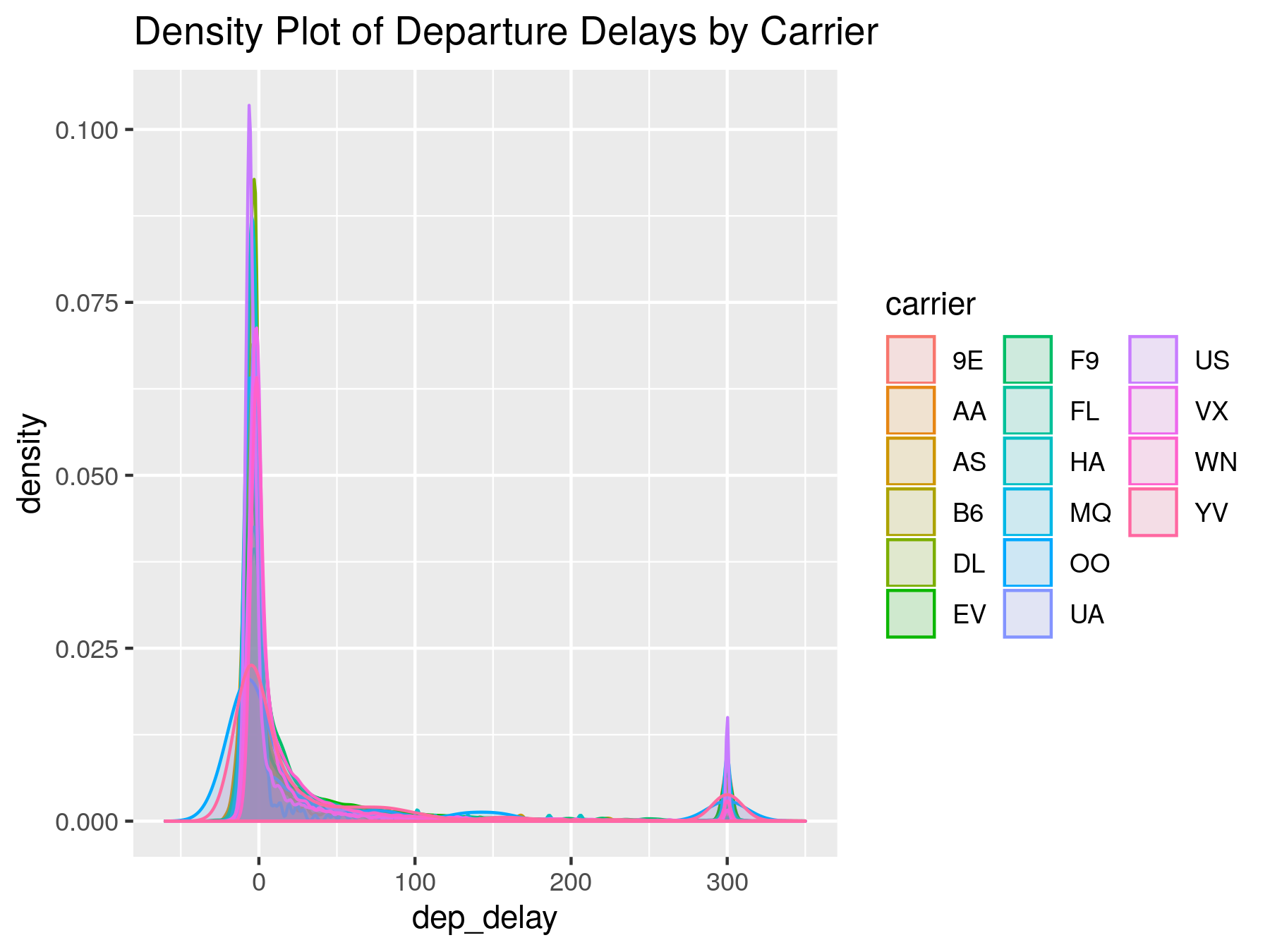
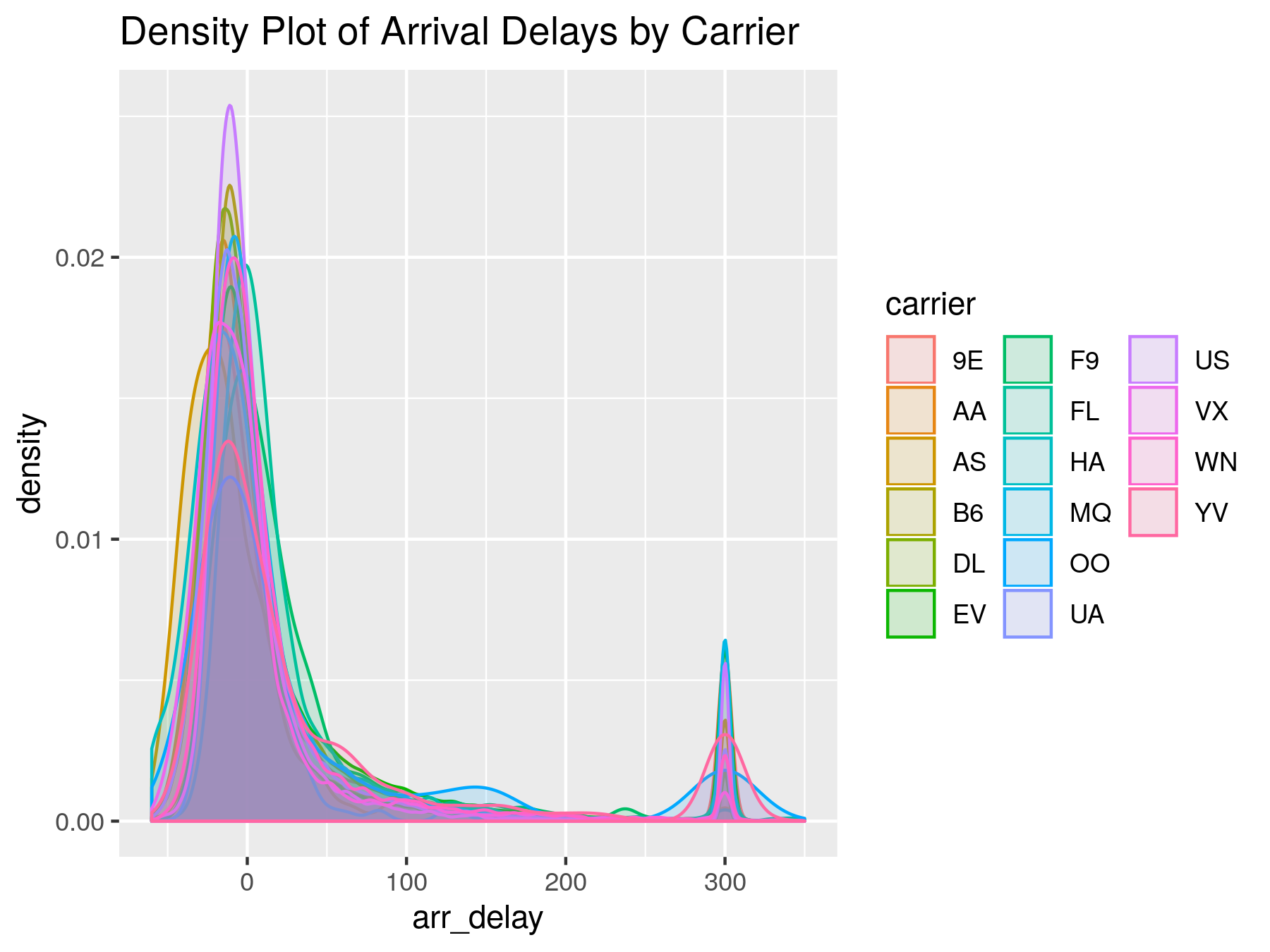
## ROC

In addition to r-squared and AIC, we also used ROC curve to evaluating and comparing the predictive models. ROC curve is a ratio curve between the TPR (sensitivity) and FPR (1-specificity)[[16]](#footnote-16). It tells us how well the model can classify the data into two classes. AUC represents the area under the curve; the higher the AUC, the better the model is.

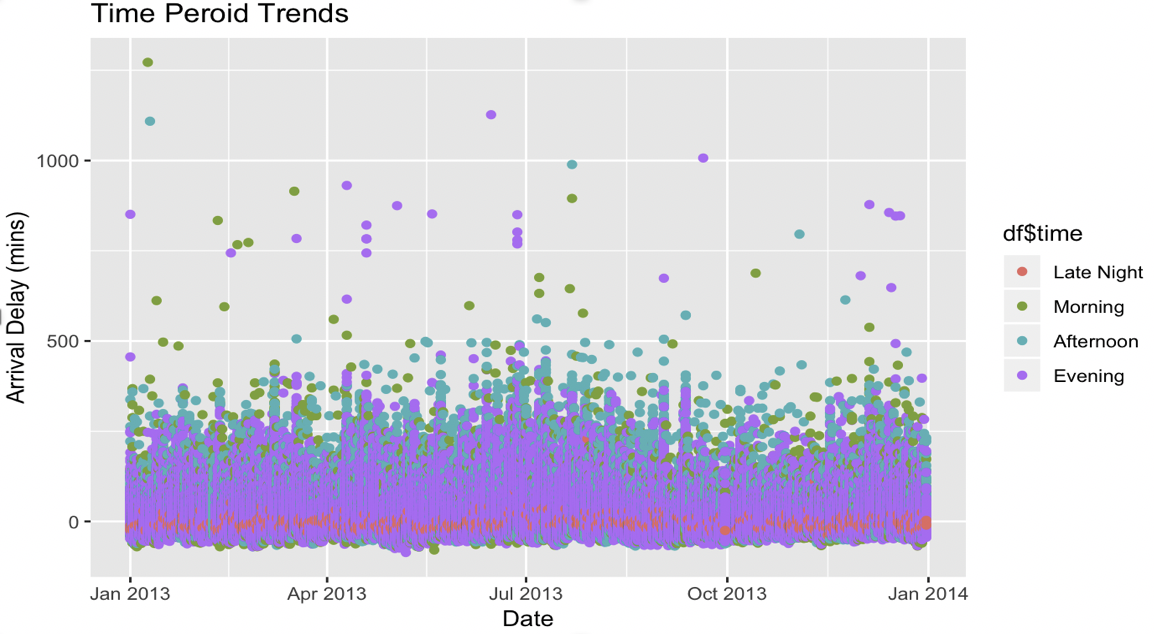
# Results

## Exploratory Analysis

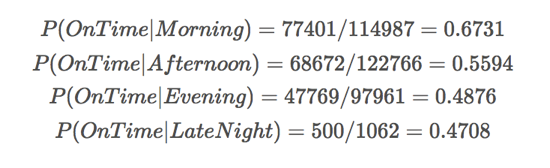
The distribution of delays by carrier shows that the center of density for the flights appears to be around the on time arrival and departure points as seen in figures 4 and 5. As discussed in the methods section there is a spike at 5 hours for delays because of the recorded missing values. The spread of the arrival delays is more than the departure delays because airplanes have the chance to make up time in the air.



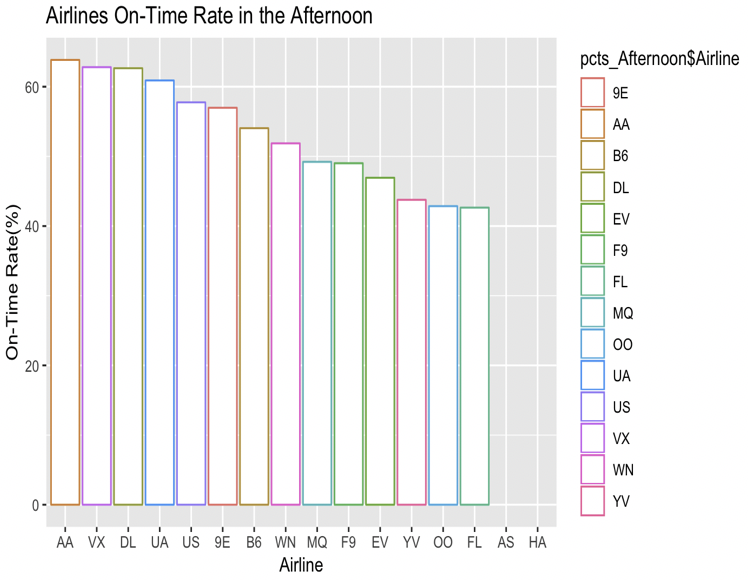
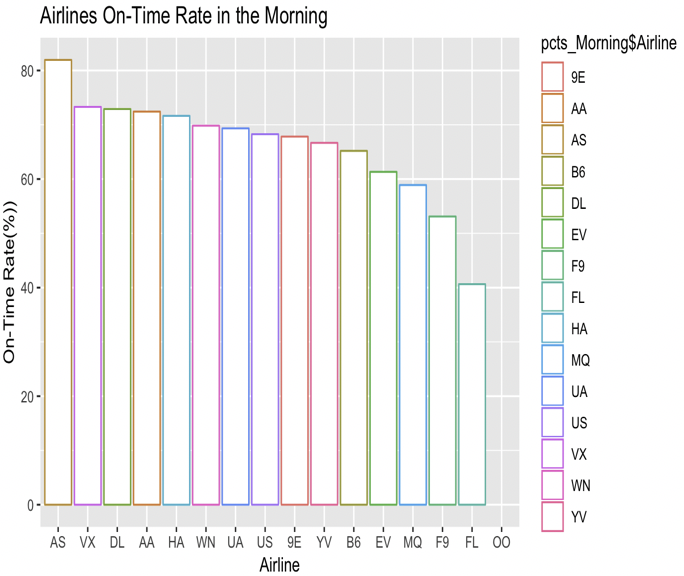
Below is the distribution of arrival delay in minutes for each of time periods on every day in 2013. As the chart displaying, the delay time ranges from negative to above 1000 minutes, the majority of delay time distribution is below 500 minutes. In this distribution, we could not clearly tell if there is a relationship between on-time rate and time periods of a day. The results of Chi-squared time for Independence to test the relationship of time periods of a day and on-time rate of flights will be discussed in later session in this paper.

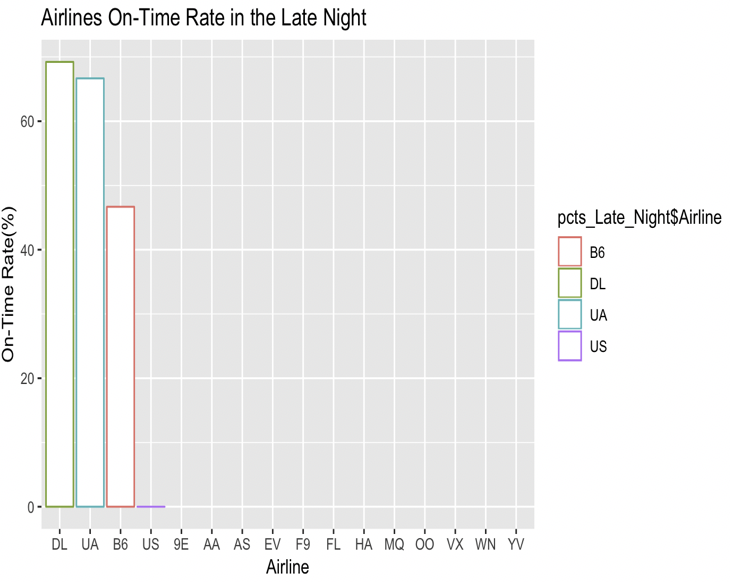
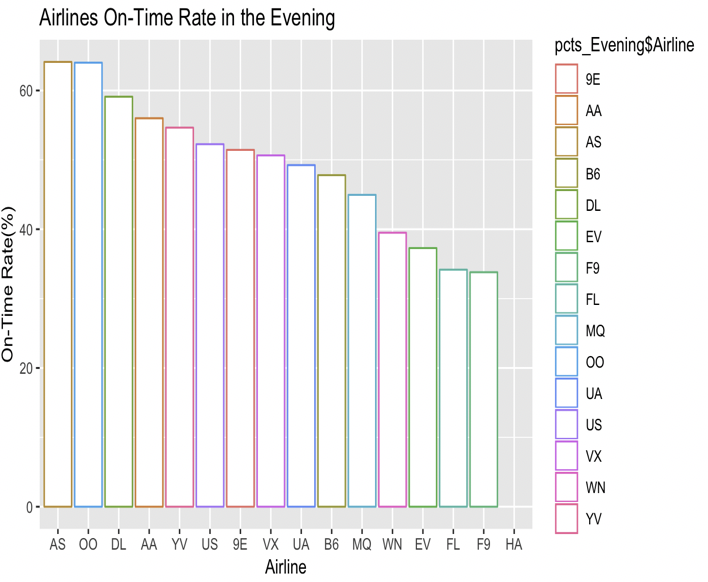


The conditional probabilities of the on-time rate of flights for each time period are shown as below. The highest on-time rate of flights is in the morning with the probability of 63.71%. In the evening and late night, the probabilities of the on-time rate of flights are below 50%.



The distributions below are showing the conditional probabilities of the on-time rate of each airline in each time periods during the day. As shown in the graphs, airlines have the best on-time performance, the majority of them have on-time rate between 60% to 70%, the probabilities range from 40% to above 80%. The second best on-time performance of flights is in the afternoon, the probabilities range from above 40% to above 60%. In the evening, half of airlines’ on-time rate are below the 50%, the probabilities range from above 30% to above 60%. In the late night, there are only four airlines have flights. Two of them have the on-time rate above 65% which is a very good on-time rate compare to others. 0ne of them has the on-time rate above 45%. The last one has 0% on-time rate, this is the reason why the average on-time rate in the evening is below 50%.





Then the project is going to focus on origin airport variable to make some exploratory analysis. Though the difference on departure delay among these three airports is not very visible from the graph. We can see from the summary, the mean delay from JFK is the shortest in these three airports and LGA is the longest.

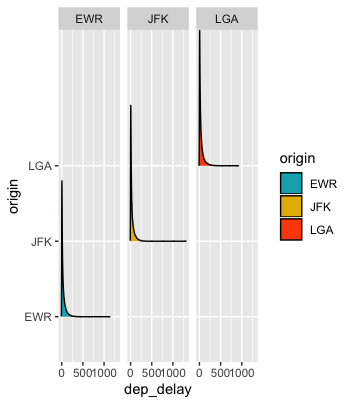
## 

## 

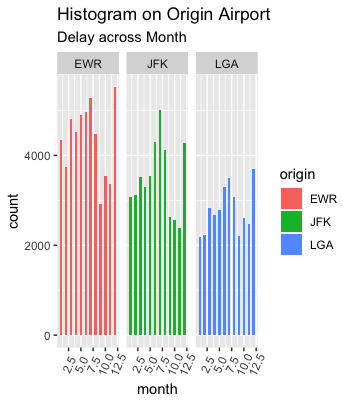
## 

## 

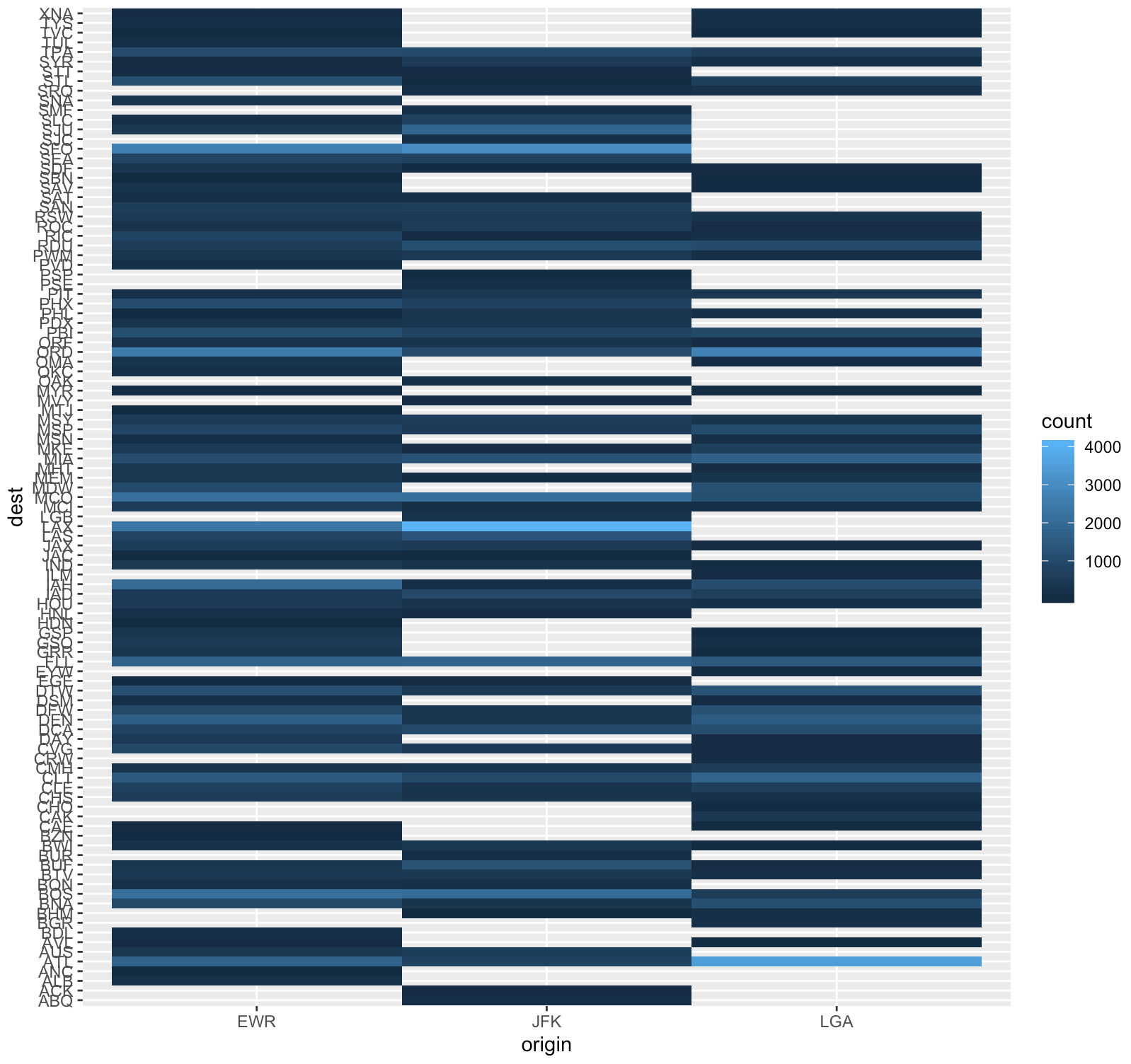
The distribution of the delay among these three airports are quite similar according to the frequency graph. So the project is going to dive deeper into the combination of other components.



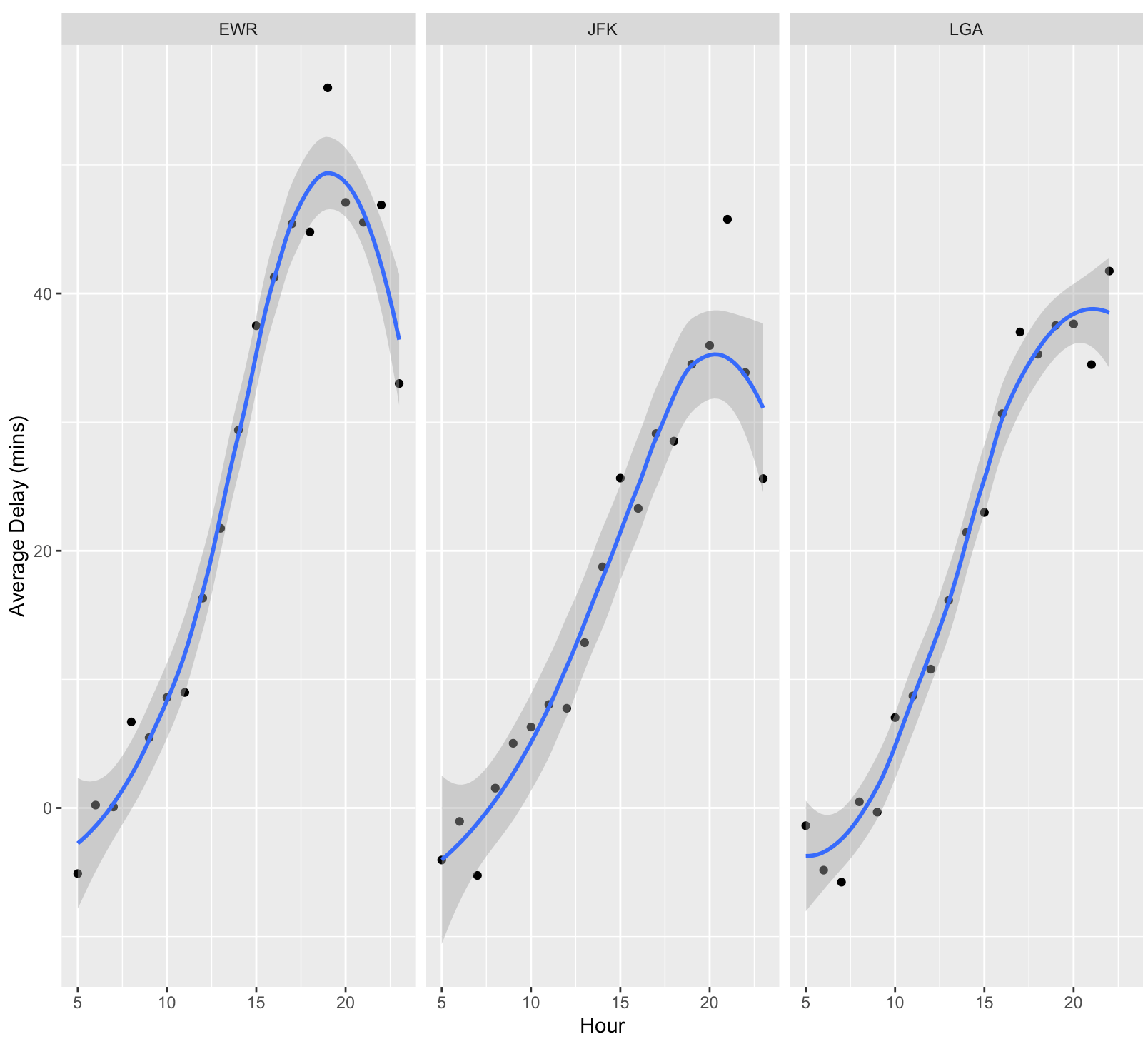
The first variable that the project dives into is month variable. We can see that the distribution of delay on month of these three airports are not identical but do share some similarities. For example, there aren’t many delays in September in all of these 3 airports. And the delay of December and July are usually heavy. The delay of January in LGA is the lowest among the three.



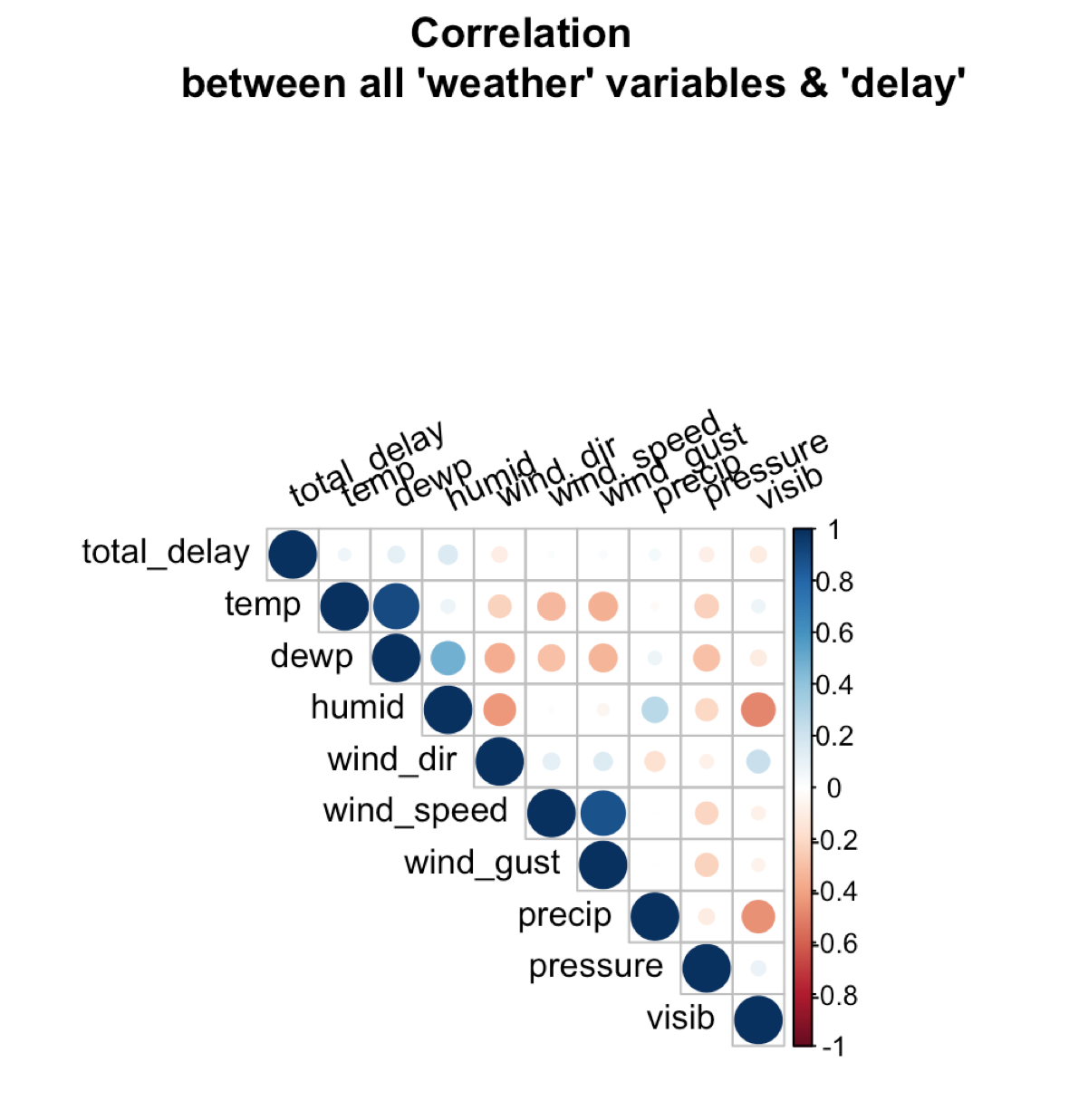
The second variable the project dive into is destination variable. By looking at the heatmap of the delay among the three origin airports the project is discussing and all the destinations, there are some noticeable suggestions. If a passenger wants to travel to the ORD airport in Chicago, he or she probably wants to depart at JFK to avoid a higher probability of delay. If a passenger wants to travel to the LAX airport in Los Angeles, he or she probably wants to depart at EWR to avoid a higher probability of delay.



The last variable that the project dives into is hour. As readers can see from the graph the distributions of the three origin airports along hour are quite similar. With the delay time increasing from 5am to 8pm and arrive at maximum delay at 8pm then start to decrease.



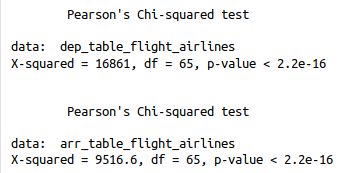
## Correlations



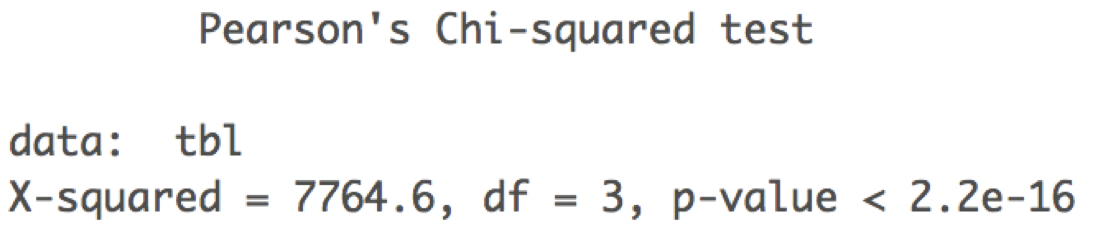
Wind direction and wind speed do not have any correlation. Thus our predictive models excluded these two variables. Furthermore, there is a strong correlation between temperature and humid so our models made sure to not include both of the variables.

## Chi-Squared Tests

The chi-squared tests were performed to test the relationship of airline carrier on both arrival and departure delays. Due to lack of data airline carriers OO and HA were dropped from the test data. The p-value result for both arrival and departure delays was below 0.01. This means that at the 5% significance level we can reject the null hypothesis that there is no relationship between airline carriers and departure and arrival delays. There is sufficient evidence to conclude that airline carriers do differ in their departure and arrival delays.



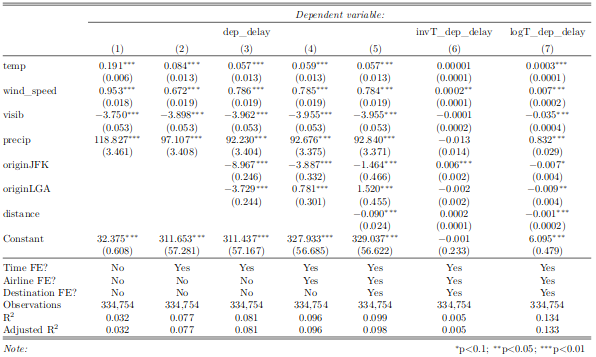
For the Chi-squared test for Independence on time periods of a day and on-time rate of flights, below is the result for this test. The p-value is very small and is smaller than the significant value 0.05, so we can reject the null hypothesis. Thus, there is enough evidence to conclude that there is a significant relationship between on time rate and daytime periods.



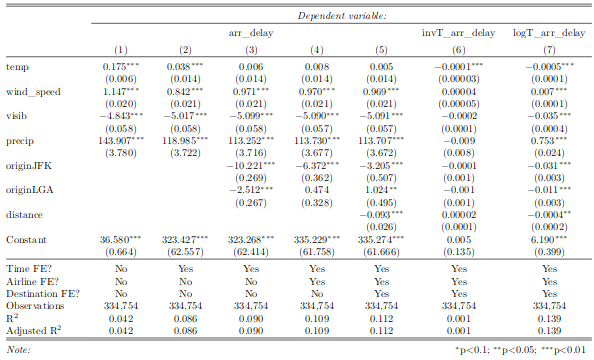
## Multiple Linear Regression

The linear regression framework was performed on both departure and arrival delays. Descriptions of the results will refer to both delay types since results are similar, unless otherwise noted. Model (1) is the effect of weather data on delays, which results in all statistically significant variables. Higher temperatures, wind speed, and precipitation all cause greater delays while lower visibility also contributes to greater delays. The R-squared value is rather low at only 0.032 for departure delays and 0.042 for arrival delays. Model (2) Includes time fixed effects in addition to the weather variables. This allows for increased delays from flights later on in the day due to problems with previous flights that cascade throughout the day. The other additional time fixed effect is the month variable, which would take into account differences in the delays that occur during the holiday month of December vs October etc. The R-squared values almost double for both regressions. Model (3) builds on model (2) and adds the categorical variable origin airport, which results in a similar R-squared value. Flights departing JFK and LGA have statistically significantly lower delays than flights departing from the base group Newark (EWR). Model (4) includes airline carrier fixed effects leading to decreased coefficients and sometimes lowered significance of departing airport. The R-squared values increase by about 0.02 for both delay types. The origin airport LGA reverse sign to be worse than Newark in departure delays, but makes up time in the air to be about the same as Newark in arrival delays. Model (5) includes the destination airport and the distance to that airport into the regression. The R-squared remains the same, but the flipped LGA effect remains. Model (6) includes all the same regressors in model (5) but the dependent variable is now the modified inverse transformed delay. The R-squared is almost 0 in both cases so we conclude that the inverse transform is not helpful. Model (7) is the same as (5) and (6) but with the modified log transformed delay. All variables in the regression are significant. We find that all weather variables affect the delay time, Newark is worst airport to leave from, and flights further away have less of a delay. Model (7) has R-squared value above 0.13, which is the strongest predictor model of the seven tested.

Departure Delay Multiple Linear Regression Results

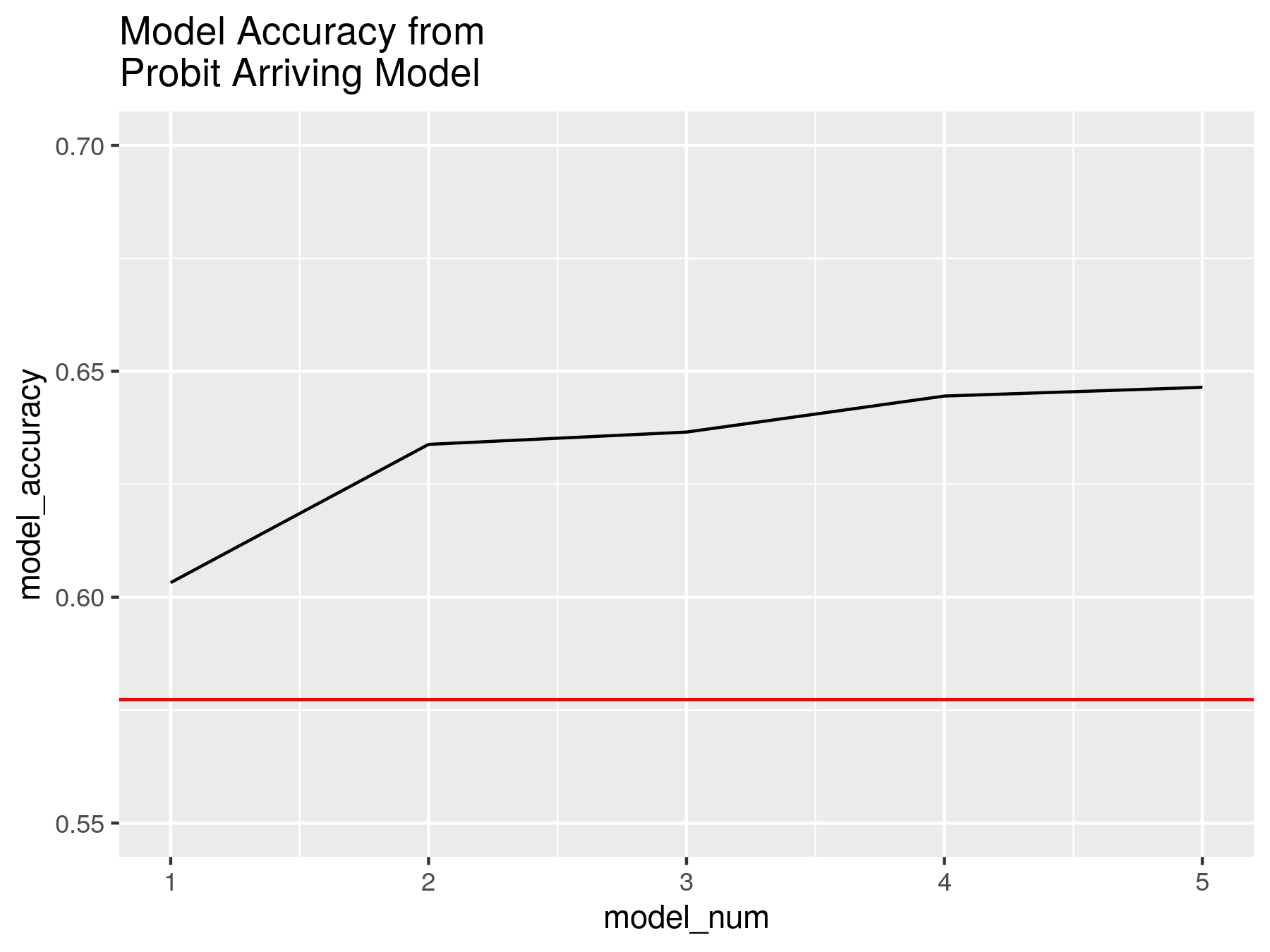
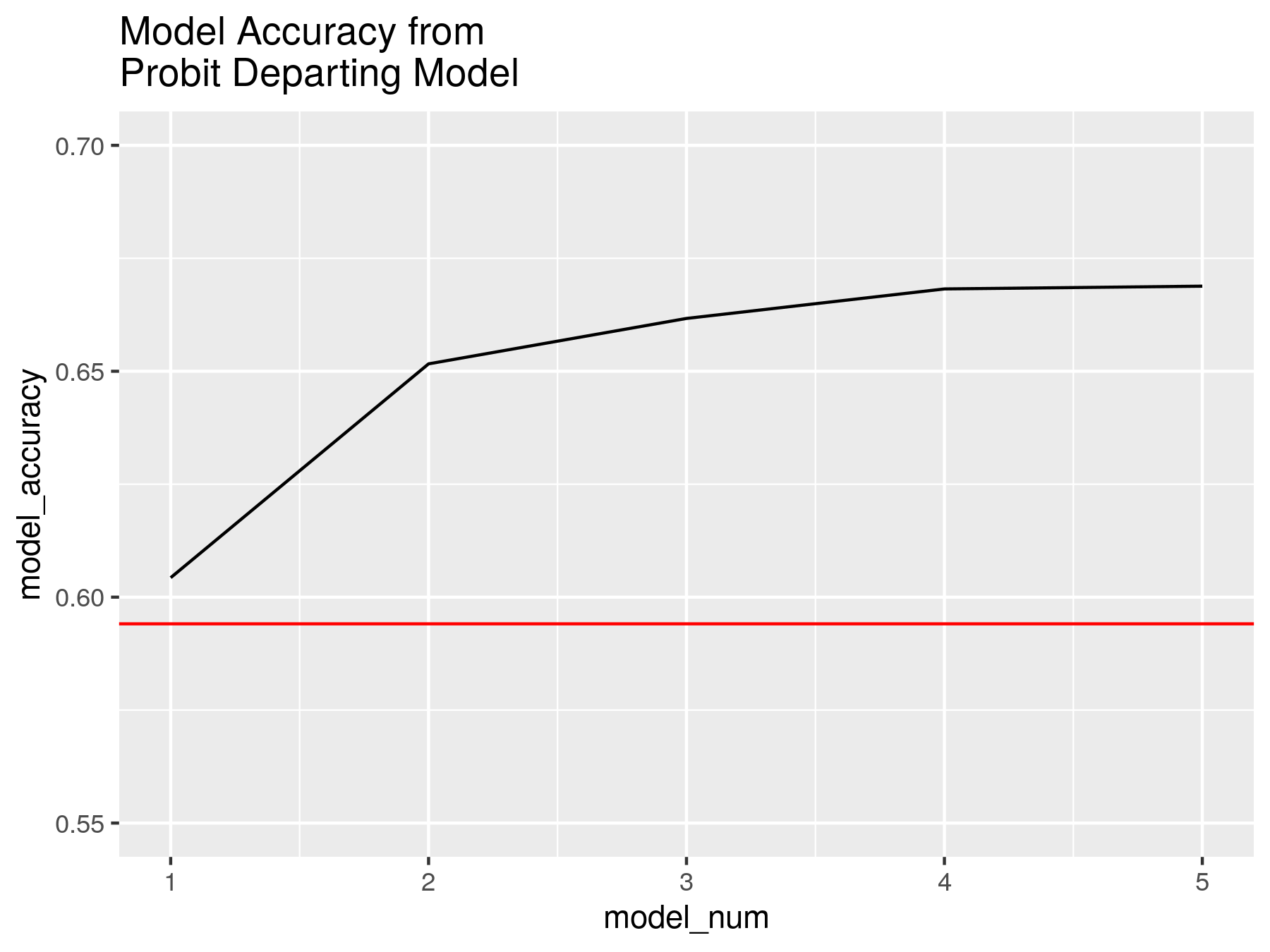


Arrival Delay Multiple Linear Regression Results



## Probit Model

Due to the limitation of only getting a 0.13 R-squared on the MLR analysis. Converting delays into binary variables allowed for further analysis and prediction. The models used in the probit regression are identical to the first five models from the MLR analysis. The red line on the graphs show the percentage of flights that have no delay. This is the baseline for the model because if the model were to just pick no delay every time it would have an accuracy at the red line. Improvements off of that constant are shown as the black line. The best probit model for both departing and arriving delays is model (5). Though much of the prediction accuracy increase is due to the addition of time fixed effects in model (2). Model (5) once again performed the best out of all the models both in terms of accuracy and also in terms of AIC.



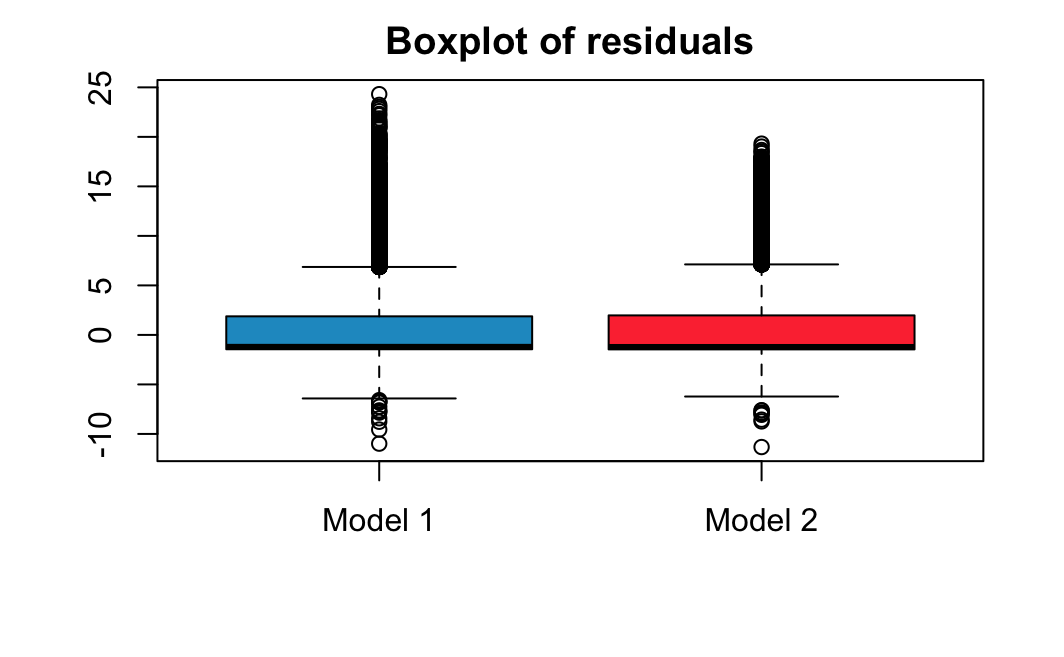
## 

## Logistic Regression Model

For the prediction that performed by the Logistic Regression method and Naive Bayes method, the departure delay and arrival delay shows the very similar results, thus we will only present and discuss the departure delay here.

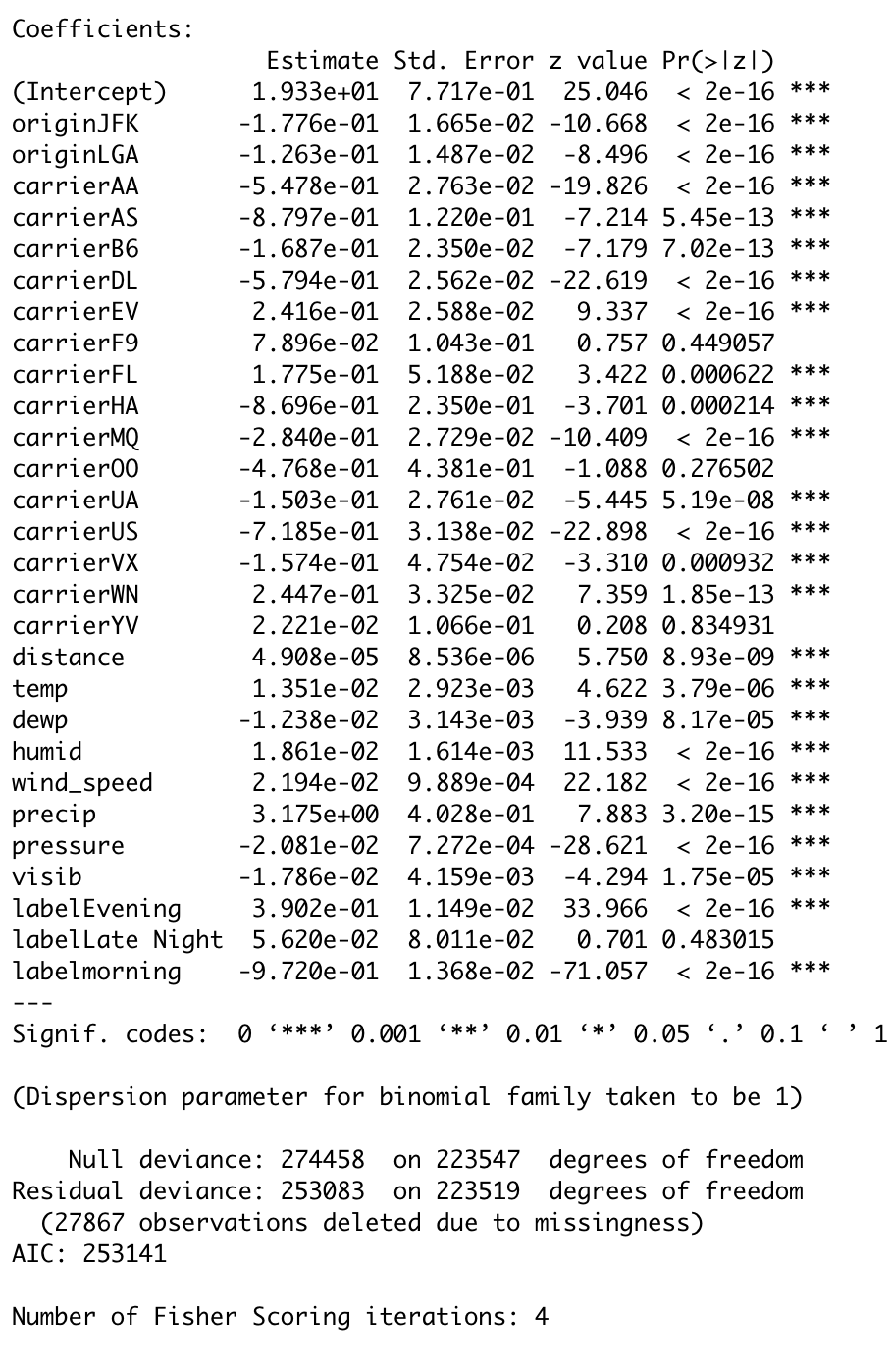
Followed by the analysis we did above. We firstly selected the highly correlated features as the regression variables - “distance, departure airport, carrier, time period, wind\_speed, precipitation, pressure, visibility”. Then we apply the logistic regression. The result shows: McFadden's R squared of regression is 0.19, which indicates that only 19.8% of improvement over the intercept model offered by the regression model; and accuracy of the model is 71.5% which is only approximately 10% higher than the random selection probability. Obviously The logistic regression doesn’t fit the data well. Since the logistic regression is sensitive to the correlated variables. We decided to introduce “Month” variables for improving the models. We marked the model with fewer variables as the model 1 and the model with more variables as model 2. The result comparison between two logistic regression models is presented below:

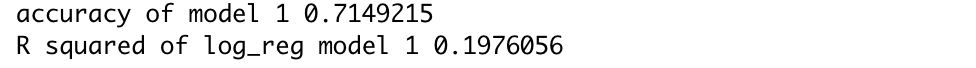
|  |  |  |
| --- | --- | --- |
|  | Model 1 | Model 2 |
| AIC | 253141 | 249750 |
| Accuracy | 71.5% | 72.2% |
| R Squared | 0.19 | 0.21 |



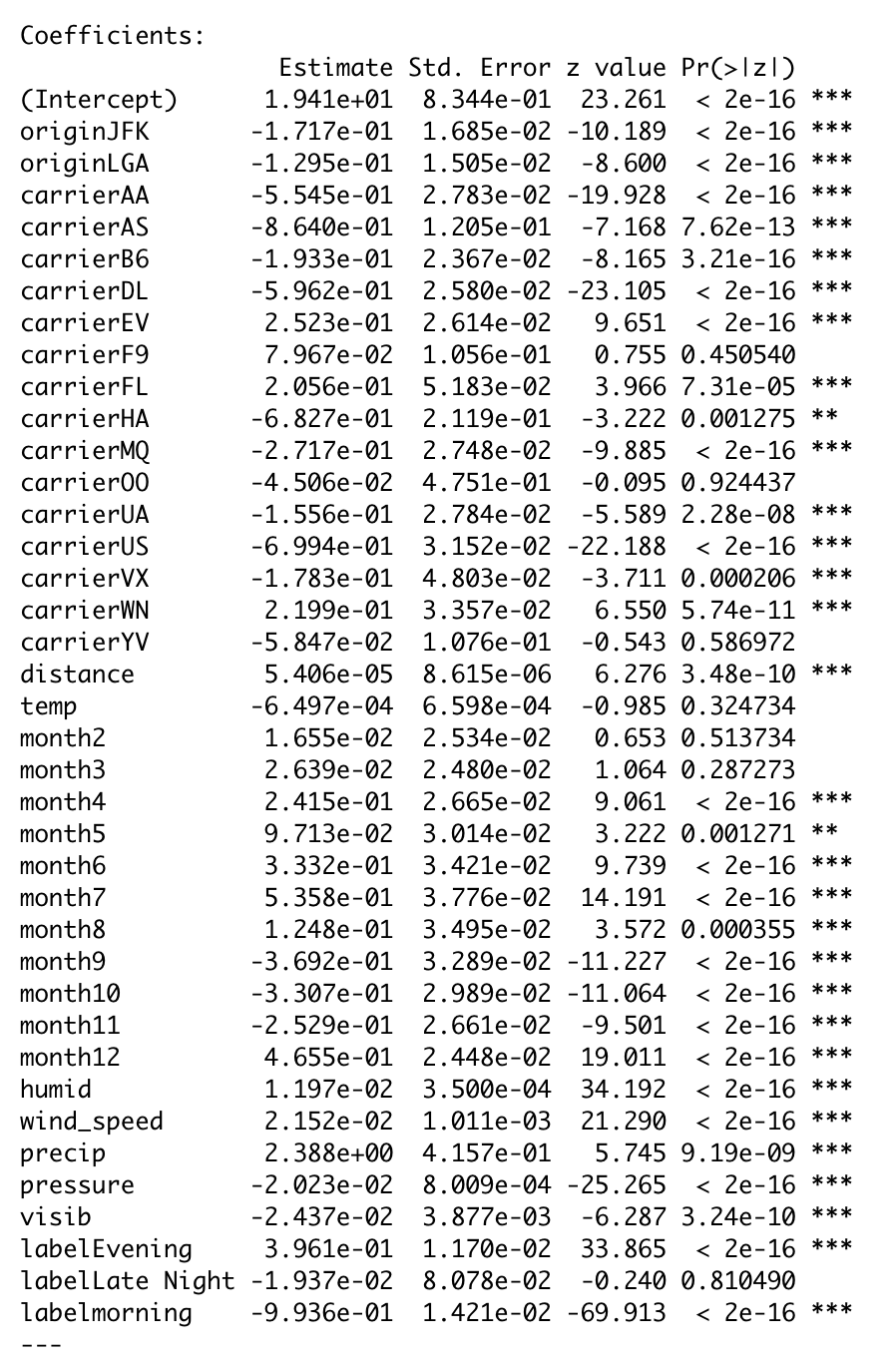
According to the table, we found that model 2 has smaller AIC, higher accuracy and larger R squared, and the residuals plot shows that model 2 holds smaller outlier range and is more concentrated to the 0. All these evidence imply model 2 has better quality, compared to model1. However, the difference between the two models is extremely small. We still can not conclude that the logistic regression can give us a good prediction of whether the flight will be delayed.

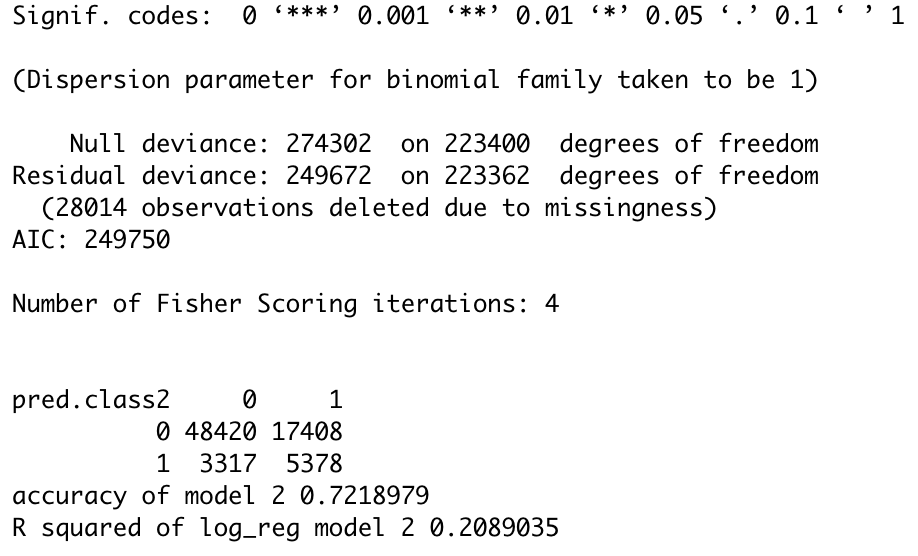
Departure Delay Logistic Regression Model 1 Results:





Departure Delay Logistic Regression Model 2 Results:

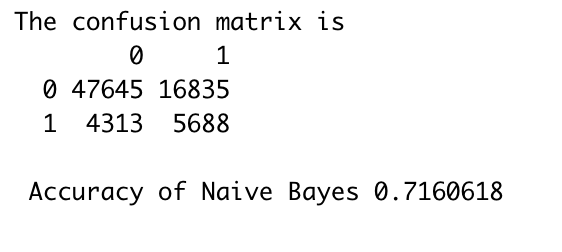




## Naive Bayes

In order to obtain better results, we also try the Naive Bayes method to predict the likelihood of flight delay. We used the same variables as the logistic regression model 2. However, the accuracy of the Naive Bayes method is 71.6%, almost no difference from the logistic regression models.

Departure Delay Naive Bayes Results:



Then we conducted the Receiver Operating Characteristic (ROC) curve to compare the probit model(the orange line), logistic regression model(the blue line) and Naive Bayes model(the violet line). The black line is the diagonal reference line, shows the random guessing result. Apparently, the area under the logistic regression curves is largest among the areas under three curves, which implies that the logistic regressions have relatively better predictions. But all three curves are close to the diagonal reference line, and the AUCs of three curves are close to 0.5. None of the prediction can give us an accurate prediction of flight delay with given variables.

The ROC plot of three prediction models:



## Discussion

Based on all the predictions model we built, we can conclude that we fail to make an accurate prediction, given the variables “distance, departure airport, carrier, time period, wind\_speed, precipitation, pressure, visibility and month”. However, according to the previous analysis we did, these variables do have an effect on the punctuality of the flight. We believe that there must be some other factors that have a greater influence on flight delays than the variables we use.

Then we did some research. We found that the variables that we used only accounted for a small part of the flight delay factor. The Federal Aviation Administration (FAA) states on their website that only 3% of delayed flights are caused by weather conditions. The air traffic control (ATC) restrictions should be responsible for at least 30% of delayed flights. And the other 25% of delayed flights are caused by the air carrier delays, for example, waiting for connecting passengers cargo or bags, loading cargo and bags, boarding passenger, aircraft preparation and so on. In addition, there are also some unexpected factors that can impact the flight delays, such as airline staff's strikes, bird strikes, problems with the coffee machine, or criminal on board. In fact most of these variables are highly random and unpredictable. This greatly reduces the possibility of obtaining the accurate prediction.

# Conclusion

In this paper, we have performed analysis of factors, time periods, origin airports, weather, and carries, which would affect flights delay. In the analysis we have made the following results. The flights in the morning have the highest on-time rate. The origin airport LGA has the highest delay rate and the origin airport EWR has the lowest delay rate. The different carries have statistically different delays. For weather analysis, we find that the wind direction and wind speed do not have any correlation with delays. We have performed further study and research on the weather conditions and delays, as we discussed in the discussion section, the weather condition has a very small portion of effects to flight delays, and there are other factors weighed much more to affect flight delays.

In the prediction, we use prediction models of Probit Regression, Logistic Regression, and Naïve Bayes. The Logistic Regression model has the highest accuracy of 72.8%. As we mentioned in the discussion section, other uncontrolled factors are significantly reducing the accuracy of the prediction model, the factors we have analyzed have limited effects to fight delays.

1. <https://www.airhelp.com/en/blog/2018-record-year-flight-delays/> [↑](#footnote-ref-1)
2. <https://www.bts.dot.gov/annual-passengers-all-us-scheduled-airline-flights-domestic-international-and-foreign-airline> [↑](#footnote-ref-2)
3. <https://www.bts.dot.gov/newsroom/estimated-october-2019-us-airline-traffic-data> [↑](#footnote-ref-3)
4. <https://www.airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/> [↑](#footnote-ref-4)
5. <https://www.airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/> [↑](#footnote-ref-5)
6. Weiwei Wu, Cheng-Lung Wu, Tao Feng, Haoyu Zhang, and Shuping Qiu, “Comparative Analysis on Propagation Effects of Flight Delays: A Case Study of China Airlines,” Journal of Advanced Transportation, vol. 2018, Article ID 5236798, 10 pages, 2018. https://doi.org/10.1155/2018/5236798. [↑](#footnote-ref-6)
7. <https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf> [↑](#footnote-ref-7)
8. <https://christophm.github.io/interpretable-ml-book/logistic.html#theory>. [↑](#footnote-ref-8)
9. <https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf> [↑](#footnote-ref-9)
10. <https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf> [↑](#footnote-ref-10)
11. <https://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/> [↑](#footnote-ref-11)
12. <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/> [↑](#footnote-ref-12)
13. <https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf> [↑](#footnote-ref-13)
14. <https://uc-r.github.io/naive_bayes> [↑](#footnote-ref-14)
15. <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/> [↑](#footnote-ref-15)
16. <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5> [↑](#footnote-ref-16)