Predicting and Analyzing Causes of Delay for Flight Departures

Shaun Tran

May 6, 2023

Abstract

Paired with a growing population is a demand for travel. In my study, I closely examined airline departure times by comparing their scheduled and actual departure times. I considered a delay to be any difference of 15 minutes or more and recorded the cause of the delay. To ensure a fair assessment of each airline's performance, I collected data from a single airport every day for 10 years. By applying the knowledge I acquired in class, I used R to employ linear regression, trained a mathematical model and optimized it with least squares to create training and testing sets. By utilizing mathematical analysis through T values, P values, and R-squared values, my results offer valuable insights into airline performance regarding flight departures. My model can be used to predict future performance and analyze past data to identify the most common causes of delay in departure statistics. A valuable resource that can help both airline companies and consumers. This is proven when I practiced my model on departing flights from Delta Airlines at the San Francisco International Airport from January 2013 until 2023, and recognized that flight carrier delay and late arriving aircraft were the biggest causes of delays, suggesting that it is well within an airline's control to improve their on-time performance.

1. Introduction **1**
a. Literature Review 1
b. Motivation 1
2. Methods **4**
a. Approach to the Mathematical Model 4
b. Multiple Linear Regression Model 6
c. Least Square Method 8
3. Analysis and Results **9**
a. Mathematical analysis 11
4. Discussion/Conclusion **19**
5. References **21**
6. Code for my Mathematical Model **23**

1. Introduction

a. Literature Review

There is a growing interest in understanding and quantifying flight delays, as evidenced by the increasing number of studies on this topic. These studies have utilized various statistical methods to analyze arrival delays at airports and compare the performance of airlines and airports. For example, one study analyzed ground delays and air holding at Entebbe International Airport over a five-year period and generated daily probabilities for aircraft departure and arrival delays(Wesonga, Nabugoomu, & Jehopio, 2012). This study established a method for comparing both mean delay and extreme events among airlines and airports, identifying a power-law decay of large delays. In other words, these studies examined the factors contributing to flight delays in 2016 using data from the Bureau of Transportation Statistics. Through reading other reports, I pinpointed that researchers employed machine learning techniques and statistical models to create a predictive modeling engine capable of identifying delays in advance. This indicated that various statistical methods were utilized in order to predict future data, and I found this intriguing. Such research offers valuable insights into the causes of flight delays and serves as inspiration for further investigation in this area. With this in mind, my team had a solid foundation with a scope of what I wanted to pursue. I wanted to utilize the knowledge I acquired in class and apply that to such a relevant issue. Through conducting this literature review, I had the opportunity to recognize that various reports covered this relevant issue in their own approach. In hopes that I would be able to provide further insight into the underlying issues behind flight delays, I found myself growing with curiosity on how I would approach my model.

b. Motivation

When coming up with my model, I wanted to examine a growing issue and how it personally relates to us. So, I decided to focus on travel as we have all experienced the frustration and inconvenience caused by a flight delay. In recent years, the global population growth has led to an unprecedented demand for air travel for various purposes such as business, leisure, and family. This increase in demand necessitates the scheduling of more flights, raising concerns about the ability of airline companies to maintain on-time performance. Flight delays can negatively impact customer loyalty.

Initially, I aimed to model the primary concerns of frequent air travelers by analyzing the air travel consumer report, which details the most common consumer complaints for each airline company such as lost luggage, refunds, etc. However, this approach only allowed me to analyze said data as a whole and identify the most significant factors on a yearly basis. Specifically, I used this approach with Markov Chains, but weren't able to define transition matrices that would portray information that would be valuable as time passes. For example, the probability of a number of complaints for a specific airline based on a certain amount of time. I instead wanted to be able to have a more detailed approach on a specific issue so I shifted my focus to one of the leading causes of consumer complaints - flight delays. That is why I instead set my sights on a mathematical model that would be done through linear regression. Once I examined closely at the statistics regarding flight delays in recent years, I distinguished that there was merely a 79.96% statistic of on-time flights for major U.S. carriers in November 2022(Bureau of Transportation). This data is publicly available information from a government resource, the bureau of transportation, and this is what certainly propelled my motivation for building my model. My interest was piqued and I began to get curious about the underlying patterns and trends. Why are flights not departing on time? What's the biggest cause and if so, is there a relationship? How could this be addressed/improved?

As I was consuming information of detailed statistics on departures, I aimed towards constructing a mathematical model that would offer valuable insights to both airline companies and consumers on their departure performance. By evaluating performance of airline companies, this indicates that they would gain the ability to discern the factors that impact their on-time performance. For example, this would allow a company to be able to justify investing more of their time and resources into improving specific issues. To improve their evaluation on timeliness on departing flights, if their largest contributing factor to delayed flights are problems revolving around the crew and scheduling flights, they would know where to focus on. Furthermore, this model's significance goes hand in hand with customers as well. If an airline's performance were to be improved, then customers would have a more confident choice in being able to choose an airline that has a good performance record. Over time, this would build customer loyalty and benefit both the company and the consumer.

So in order to address this issue, I decided to examine the overall causes of delays by reporting operating carriers. By examining data from major U.S. flight carriers, there were a myriad of reasoning behind each delay. Air carrier delay would be due to circumstances within the airline's control and this could be related to maintenance or crew problems, etc. Extreme weather delay is due to significant weather conditions that in the judgment of the carrier, delays the operation of the flight. National Aviation System Delay is caused by non-extreme weather conditions, such as airport operations, heavy air traffic volume, etc. Security delay is caused by security issues such as evacuation situations, security breaches, long lines at screening areas,etc. Lastly, late arriving aircraft delay which suggests that a previous flight arrived late, forcing the present flight to depart late. This is why the assumptions for my model will be that I will be predicting flight departure times, based on these causes of delays as my only predictors for my model. To make the assumptions for my model more fair, I utilized data on the same airport. By choosing to only take into account flights that departed from the San Francisco International Airport, for every day from January 2013 to January 2023, I would be able to determine weather delays as a more fair assessment. I also chose San Francisco International Airport due to how it has abundant flights departing everyday, so I could input plenty of data.

Furthermore, in order to predict departure data based on historical flight performance, I wanted to focus on evaluating an air carrier that performed well in departing flights on-time. This is why I wanted to concentrate my model around Delta Airlines. Based on the data from the Bureau of Transportation from 2022, Delta Airlines had one of the highest total records of flights in the U.S. paired with consistently departing on time, at an impressive rate of 89.91%(Bureau of Transportation). Once my model is utilized, it would be extremely significant as it would have the ability to accurately assess the performance of Delta Airlines, predict future performance, and identify the biggest factors that contribute to flight delays. By identifying these factors, I can recognize if Delta Airlines has room for even further improvement. Hence, my mathematical model holds considerable importance for both the airline and its customers in improving the flying experience.

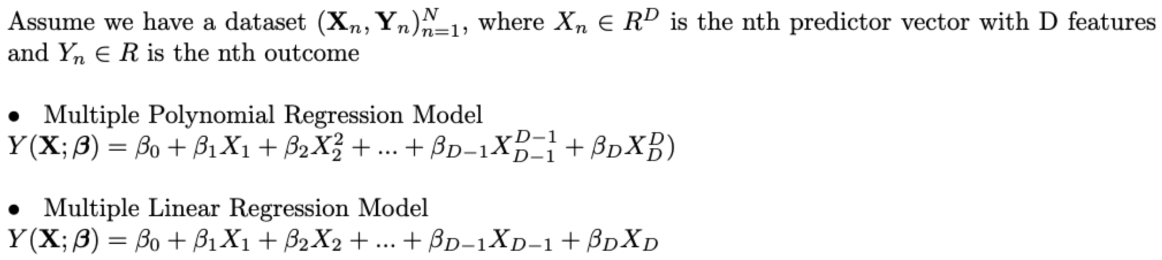
2. Methods

a. Approach to the Mathematical Model

When I decided which types of mathematical model to apply, I expected that the model could capture the pattern of the dataset and explain the relationship between input and output. Moreover, according to the note from CSE 176 (Introduction to Machine Learning) chapter 1, such a mathematical model needed the data to predict outcome and somehow gain knowledge. In other words, the model was a compressed version of the dataset and I could extract insight on flight delay while approximating the data with high performance (Carreira-Perpiñán, 2019, pg.1). Machine learning models would be adequate to be my choice of model as I had enough data for the model to make inference about flight delay. Machine learning models could be broken into supervised, unsupervised and semi-supervised learning models, where it required both inputs and labels data, only required input data and it required both input data and labels data for some data point, respectively. I was feeding my both input and label data into the model with all the data points; therefore, I chose supervised learning methods. Supervised learning models were further broken down into classification and regression models, where the differences between were on the label/outcome and the purpose of the models. Classification models were used to separate data points into different categories and outcome were qualitative while regression models are used to make predictions based on historical data and outcome are quantitative. The objective of my project was to use the supervised model to make predictions in the future time and answer my second research question - to make inference about the biggest cause and possible relationships between input and label, based on my given historical data. In addition, the label “Departure Delay” was quantitative data, and I was confident that choosing a regression model could answer my research question. There were two more considerations to choose a specific regression model. One was whether the method should be parametric or non-parametric and another one was complexity of the model according to the bias-variance tradeoff. For the parametric/non-parametric models, according to note from on CSE 176 (Introduction to Machine Learning) chapter 8, parametric model had explicit function form such as Gaussian distribution, logistic function or linear function whereas non-parametric model did not (Carreira-Perpiñán, 2019, pg. 31). I decided to use a parametric model over another because I wanted to have assumed the function form for my regression model. Choosing a parametric model came with several advantages on the complexity and dataset itself. As I had an explicit function form for my model, it would be easier to interpret my model to address my research questions and achieve my objectives. Contrary to the non-parametric model, as it was assumed to accept any function types, the training was purely based on the dataset itself. The training process became more computationally expensive. According to slides from MATH 180 (Modern Applied Statistics) lecture 3, non-parametric model suffered from the problem of "Curse of Dimensionality", which referred to the fact that the data points tended to stay far away from each other in higher dimensions, which the data points tended to be less informative (Rube, 2022). I needed much more data just to make the data points to be concentrated and prevent overfitting, which increased the memory space. However, according to CSE 176 chapter 8 note, non-parametric models would out-perform parametric models by being capable of learning complex/nonlinear relationships between input and output data (Carreira-Perpiñán, 2019, pg. 32). This disadvantage of the parametric model led me to consider the effect of bias-variance tradeoff. Based on the MATH 180 lecture, simple models tended to have high bias and underfit but it was robust to random noise, whereas flexible models could capture nonlinearity but easy to overfitting, which they included the random noise during the training and tended to have high variance (Rube, 2022). Regarding this consideration, I decided to apply multiple linear regression models. However, the thought process behind choosing this specific model would be addressed in the next part as it's more relevant to be mentioned.

b. Multiple Linear Regression Model

MATH 150 (Mathematical Modeling) lecture slides 22 introduced a general regression model which is polynomial based. However, I started off with a regression model which is linear based. In other words, I assumed that the parametric function form is linear. Therefore, I had a multiple linear regression model as my assume was that the input had more than one feature, which was the linear combination of X_n and β_n. The formula for both multiple polynomial and linear regression models was demonstrated in Figure 1. I wanted to start off with the simplest regression model to see if this model was enough to explain the relationship between predictors and response variables. If the model was not enough to predict well about the dataset, I would start to consider a more flexible model such as a polynomial regression model.



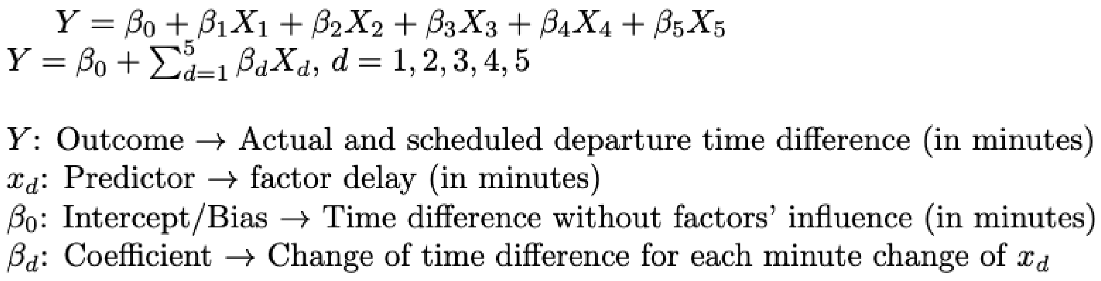
Assume we have a dataset $(\mathbf{X}_n, \mathbf{Y}_n)_{n=1}^{N}$, where $X_n \in R^D$ is the nth predictor vector with D features and $Y_n \in R$ is the nth outcome

- Multiple Polynomial Regression Model
$Y(\mathbf{X}; \boldsymbol{\beta}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + ... + \beta_{D-1} X_{D-1}^{D-1} + \beta_D X_D^D)$

- Multiple Linear Regression Model
$Y(\mathbf{X}; \boldsymbol{\beta}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{D-1} X_{D-1} + \beta_D X_D$

Figure 1: *It shows the polynomial regression model from class and linear regression model by my choice.*

More details about the model's element should be explained about the multiple linear regression model in order for me to have a solid understanding of the model and be able to interpret my results later. Figure 2 explained all the variables and parameters in the context of statistical analysis and my project on flight delay.



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

$$Y = \beta_0 + \sum_{d=1}^{5} \beta_d X_d,\ d = 1, 2, 3, 4, 5$$

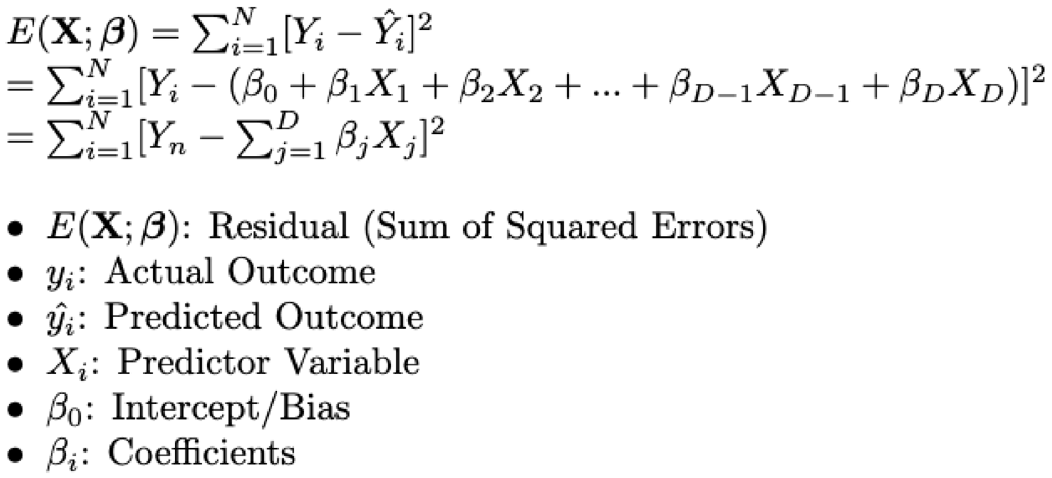
Y: Outcome → Actual and scheduled departure time difference (in minutes)
x_d: Predictor → factor delay (in minutes)
β_0: Intercept/Bias → Time difference without factors' influence (in minutes)
β_d: Coefficient → Change of time difference for each minute change of x_d

Figure 2: *It explains the meaning of each variable or parameter.*

There were two variables in the model, which were X (Independent variable) and Y (Dependent variable) and two parameters, which are β_0 (bias) and β_d (coefficients). Y is the outcome, which refers to the actual & scheduled departure time difference and unit is minutes. I defined (1) $Y = Actual\ Departure\ Time - Scheduled\ Departure\ Time$. When $Y > 0$, there is a flight delay; when $Y = 0$, the flight is on time and when $Y < 0$, the flight is early. (2) X_d is a predictor variable, which refers to factor variables in minutes. When $X_d > 0$, there is a flight delay; when $X_d = 0$, flight is on time and is early. However, X_d cannot be negative as its flight delay is already mapped to positive value, and on-time and early time mapped to zero value. (3) β_0 is the intercept of the function or bias, it refers to the time difference without influence from any factors, which is unitless. When $\beta_0 > 0$, flight still delays despite of no influence from delay factors; when $\beta_0 = 0$, if there is no influence from the delay factors, the flight is on-time, and when $\beta_0 < 0$, the airplane will depart early if there is no external influence. (4) β_d is the coefficient, which represents the change of the time difference for each minute change of X_d, in which $\beta_d \in (-\infty, \infty)$.

c. Least Square Method

When I tried to use my multiple linear regression model to make predictions for the time difference between scheduled and actual departure delay, I realized that the values of coefficients were not defined yet. According to the MATH 150 Lecture 22, applying the Least Square Method could find the set of coefficients that I needed. (Zhao, 2023). I defined the objective function $E(X; \beta)$ as the sum of the square of the difference between predicted and actual outcome (difference is what I called residual). Predicted outcome y_head was represented by the regression model. Figure 3 displayed the formula and details about the least square equation/error function. Moreover, the reason that the error squared was because I focused only on the magnitude of the error and it could be differentiated. If I could find the set of coefficients that minimize the residual, the model prediction response was going to be closer to the actual outcome response, which implied that the linear regression model improves its prediction performance.

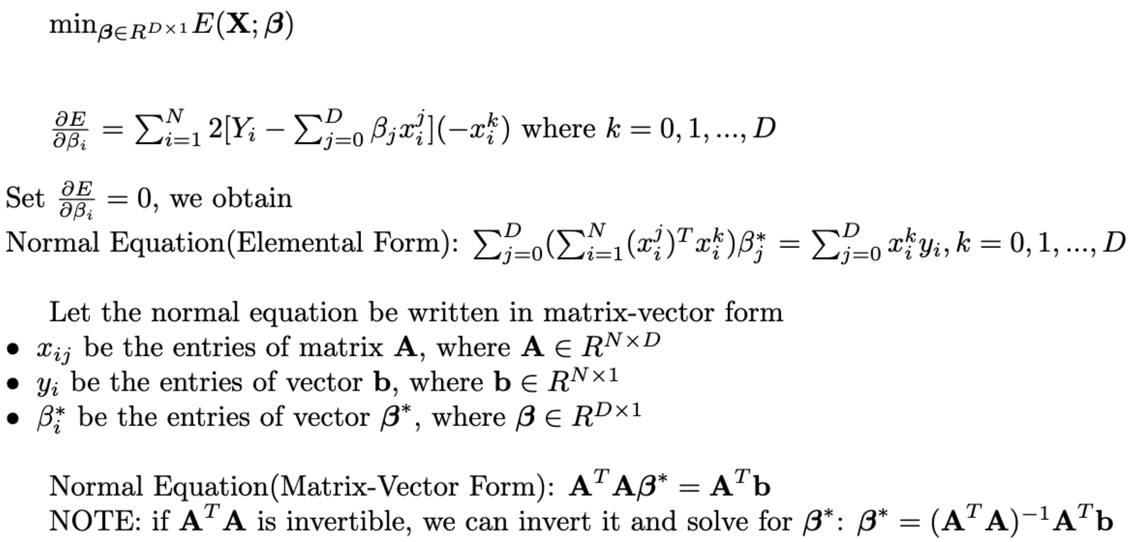


$$E(\mathbf{X};\boldsymbol{\beta}) = \sum_{i=1}^{N}[Y_i - \hat{Y_i}]^2$$
$$= \sum_{i=1}^{N}[Y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{D-1}X_{D-1} + \beta_D X_D)]^2$$
$$= \sum_{i=1}^{N}[Y_n - \sum_{j=1}^{D}\beta_j X_j]^2$$

- $E(\mathbf{X};\boldsymbol{\beta})$: Residual (Sum of Squared Errors)
- y_i: Actual Outcome
- $\hat{y_i}$: Predicted Outcome
- X_i: Predictor Variable
- β_0: Intercept/Bias
- β_i: Coefficients

Figure 3: *Least Square Formula and its details on variables and parameters.*

For the process of applying the Least Square method, Figure 4 demonstrated general steps which including defining the error function $E(X; \beta)$, solving the partial derivative $\frac{\partial E}{\partial \beta_l}$, setting it to zero and yielding the normal equation in both element and matrix-vector form.



$\min_{\boldsymbol{\beta}\in R^{D\times 1}} E(\mathbf{X};\boldsymbol{\beta})$

$\frac{\partial E}{\partial \beta_i} = \sum_{i=1}^{N} 2[Y_i - \sum_{j=0}^{D} \beta_j x_i^j](-x_i^k)$ where $k = 0, 1, ..., D$

Set $\frac{\partial E}{\partial \beta_i} = 0$, we obtain
Normal Equation(Elemental Form): $\sum_{j=0}^{D}(\sum_{i=1}^{N}(x_i^j)^T x_i^k)\beta_j^* = \sum_{j=0}^{D} x_i^k y_i, k = 0, 1, ..., D$

Let the normal equation be written in matrix-vector form

- x_{ij} be the entries of matrix $\mathbf{A}$, where $\mathbf{A} \in R^{N\times D}$
- y_i be the entries of vector $\mathbf{b}$, where $\mathbf{b} \in R^{N\times 1}$
- β_i^* be the entries of vector $\boldsymbol{\beta}^*$, where $\boldsymbol{\beta} \in R^{D\times 1}$

Normal Equation(Matrix-Vector Form): $\mathbf{A}^T\mathbf{A}\boldsymbol{\beta}^* = \mathbf{A}^T\mathbf{b}$
NOTE: if $\mathbf{A}^T\mathbf{A}$ is invertible, we can invert it and solve for $\boldsymbol{\beta}^*$: $\boldsymbol{\beta}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$

Figure 4: *The figure specified the process of Least Square Method and Normal Equation both in elemental and matrix-vector form.*

The vector of critical points β is an optimizer for which $E(X; \beta)$ reaches its optimum. However, whether this is a minimum or maximum point still remains unknown. From MATH 140 (Mathematical Optimization) lecture note, "if a function is a convex function, then the optimizer is the global minimizer." (Bhat, 2022). The theorem means that $E^*(X; \beta)$ is a minimum value and there is only unique set of βs, which gave me a unique solution. In my context, $E(X; \beta)$ is a quadratic function which is a convex function as it's twice-differentiable. The optimizer β^* is the unique minimizer.

3. Analysis and Results

Before I begin my analysis, Recall in my assumption that I am predicting the accuracy of flight departure time only using the delays I listed as predictors. I hypothesize that all the predictors are significant in my model. I used the R programming language for my analysis and I assigned 80% of my data as training and 20% as testing for my model. The reason why I chose a 80/20 split is because I don't want to over or under fit the model. Here is the graph of my prediction using all the predictors (Figure 5).

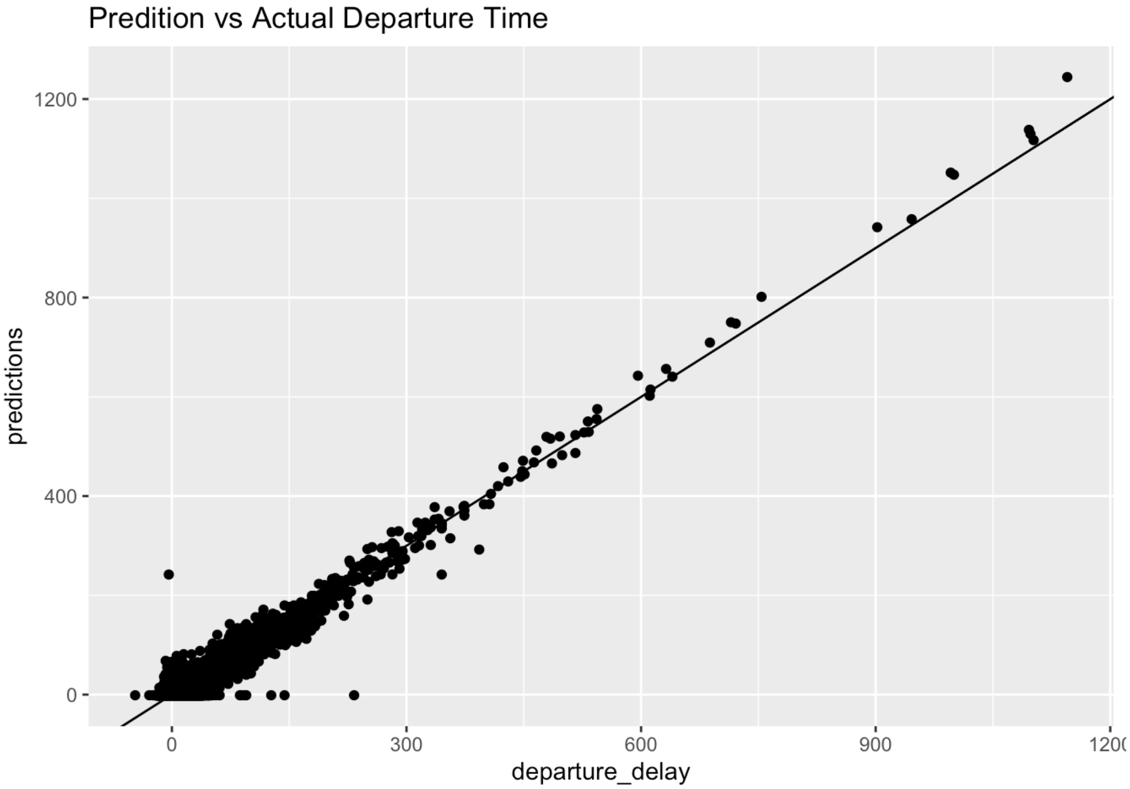
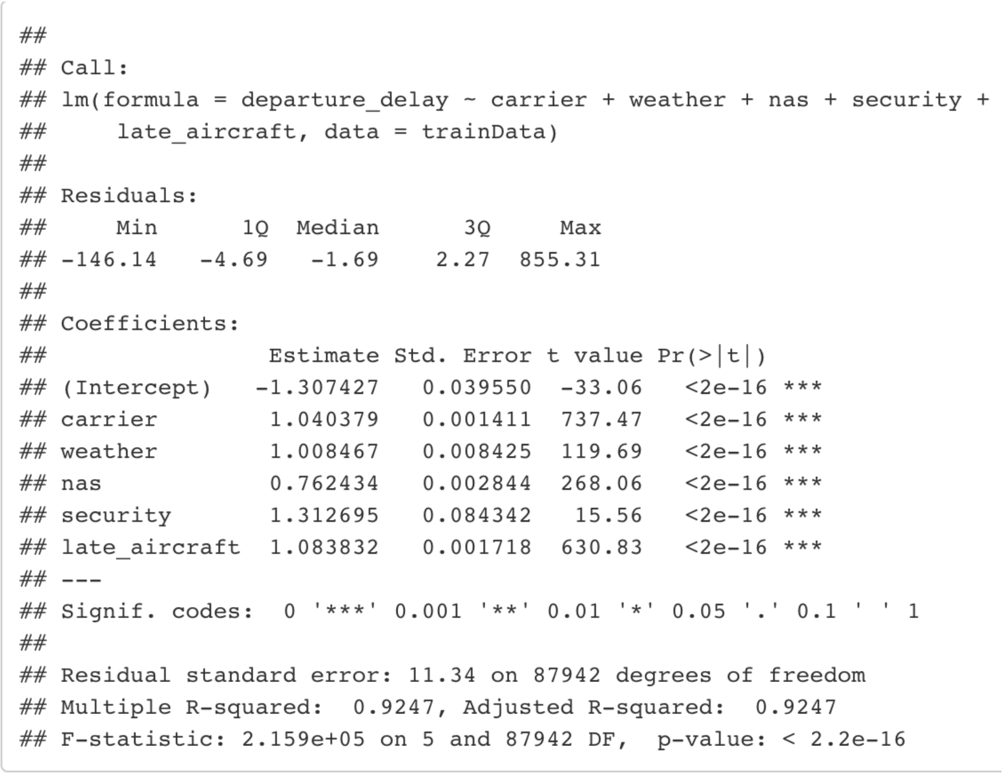


Figure 5: *Prediction Graph With Original Dataset*

Below is the summary of the linear model I created using all the predictors (Figure 6) where departure delay is my outcome variable and the five different types of delay are the predictors variables.



```
##
## Call:
## lm(formula = departure_delay ~ carrier + weather + nas + security +
##     late_aircraft, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -146.14   -4.69   -1.69    2.27  855.31
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.307427   0.039550  -33.06   <2e-16 ***
## carrier        1.040379   0.001411  737.47   <2e-16 ***
## weather        1.008467   0.008425  119.69   <2e-16 ***
## nas            0.762434   0.002844  268.06   <2e-16 ***
## security       1.312695   0.084342   15.56   <2e-16 ***
## late_aircraft  1.083832   0.001718  630.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.34 on 87942 degrees of freedom
## Multiple R-squared:  0.9247, Adjusted R-squared:  0.9247
## F-statistic: 2.159e+05 on 5 and 87942 DF,  p-value: < 2.2e-16
```

Figure 6: *Summary of Model With Original Dataset*

a. Mathematical analysis

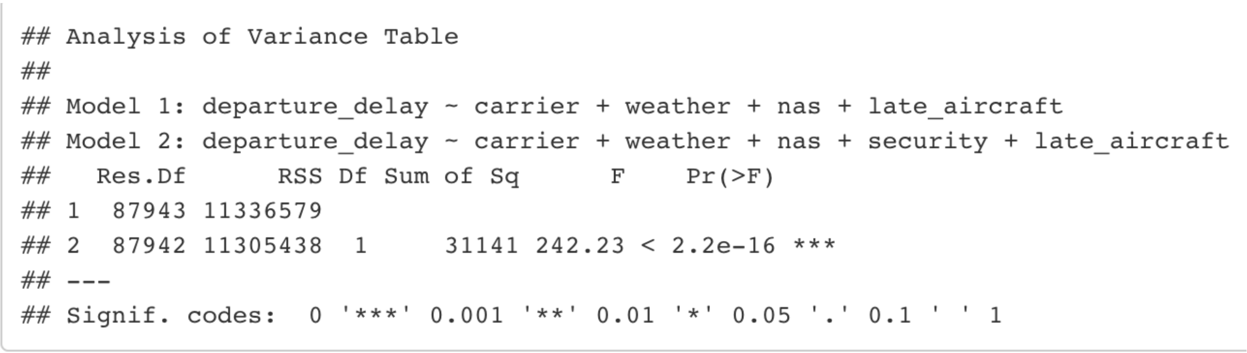
Interpreting Figure 6 from top to bottom, I can see the residuals range from -146.14 to 855.31, the first quartile is -4.69, the third quartile is 2.27 and the median is -1.69. A small Interquartile range means that the data is clustered together near the mean. And the explanation to a large min and max is that there are outliers in my dataset. The intercept is the predicted value of the response variable when all predictor variables are zero. The intercept I obtained in my model is -1.307. You may ask, does it make sense to have a negative value if I am talking about departure time? As mentioned before, a negative y value means that the plane departs earlier than scheduled. The meaning behind my intercept being negative means that when there are zero delays within my predictors, the planes leave on an average of 1.3 minutes earlier than the scheduled departure time. This could be due to the fact that planes have to fly multiple trips a day, leaving early could result in arriving early to the destination, this is beneficial because arriving earlier could avoid any delay in the next flight. One fun fact is that planes like to leave earlier in the morning because the temperature is much cooler, permitting the aircraft to depart with more payload or saving fuel(Shams,2022). The numbers under estimate are the coefficients, they represent the slope of each predictor. The standard error is a measure of the variability or uncertainty of a statistic, such as the mean or regression coefficient. The smaller the standard error, the more precise the estimate is likely to be and the standard errors of my predictors are extremely low.

In my model, I set the null hypothesis that there is no significant difference between the predictors and the response variable and the alternative hypothesis is that there is some relationship. A P-value is used to determine if I could reject my null hypothesis or not. In my model, I am setting the p value to 0.01 to be standard. I can see that there are three asterisks next to my p value, which means these predictors are statistically significant and I can reject my null hypothesis and accept my alternative hypothesis. T value is used to assess the significance of each coefficient in the model. A large absolute t-value and a low associated p-value indicate that the coefficient is statistically significant
But the T value for security is relatively low when compared with other predictors, and I conclude that the cause of this is due to the small sample size in security in my dataset.

Another interesting point I discovered when interpreting my summary is that the coefficient of NAS is less than one. Statistically speaking, this is allowed, but if I think logically, how can a one minute delay in NAS cause a forty second delay in departure time. To explain this, I have to look into how NAS is measured. And after my research, one of the factors to NAS delay is air traffic congestion, and this could occur after the plane has already departed. For example, a plane could be redirected or put on hold until conditions improve. This is considered a NAS delay but this occurs when a plane has departed. This explains why the coefficient is less than 1. And this also proves that some of the data points do not really apply to my model because of the way NAS delay was measured. If I want to use NAS as a predictor for my model, I have to find a way to separate the different types of cause in NAS delay to have a more accurate model. Ultimately, I concluded that NAS is not statistically significant, in other words, it is not an important predictor.

In general, the adjusted R-Squared value is between 0 and 1, the higher the adjusted R-squared value is, the better fit the model. The adjusted R-squared value for my model is 0.9247. Which is considered high and means that the model is a good fit.

Based on the T value in my model summary, I concluded that there are not enough data points from the security predictor in my model to make it a significant predictor. To prove this, I used the "Anova" function in R to test if adding security as an additional predictor in my model makes an impact. (Figure 7)

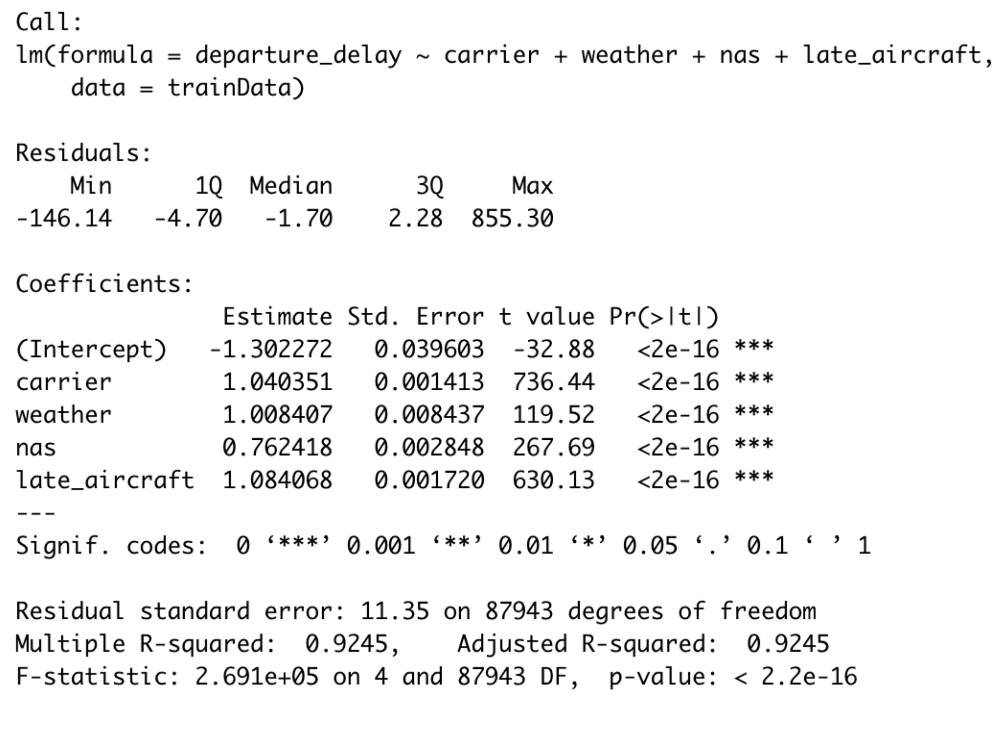


```
## Analysis of Variance Table
##
## Model 1: departure_delay ~ carrier + weather + nas + late_aircraft
## Model 2: departure_delay ~ carrier + weather + nas + security + late_aircraft
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1  87943 11336579
## 2  87942 11305438  1     31141 242.23 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7. *Analysis of Variance Table of Two Models With and Without Security*

Even though the F value is moderately high. But considering the sample size I have, I conclude that security is not an important predictor in my model with my current dataset, but if I use another dataset where there is a high amount of security delay, then it would show more significance.

Here is the result of my model without security and here is the result (Figure 8)



```
Call:
lm(formula = departure_delay ~ carrier + weather + nas + late_aircraft,
    data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max 
-146.14   -4.70   -1.70    2.28  855.30 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.302272   0.039603  -32.88   <2e-16 ***
carrier        1.040351   0.001413  736.44   <2e-16 ***
weather        1.008407   0.008437  119.52   <2e-16 ***
nas            0.762418   0.002848  267.69   <2e-16 ***
late_aircraft  1.084068   0.001720  630.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.35 on 87943 degrees of freedom
Multiple R-squared:  0.9245,	Adjusted R-squared:  0.9245 
F-statistic: 2.691e+05 on 4 and 87943 DF,  p-value: < 2.2e-16
```

Figure 8. *Summary of Model With Original Dataset Without Security*

The T value and P value compared to the initial model is almost the same, and the adjusted R-squared value is 0.9245, essentially the same, which concludes that it does not impact the model with or without security as a predictor in my model using my current dataset.

To ensure my results are correct, I plotted each predictor individually against the response variable to see their correlation. But to be more precise, I removed the 0's of each predictor from my dataset because I am only looking at the impact when there is a delay and here are the results.

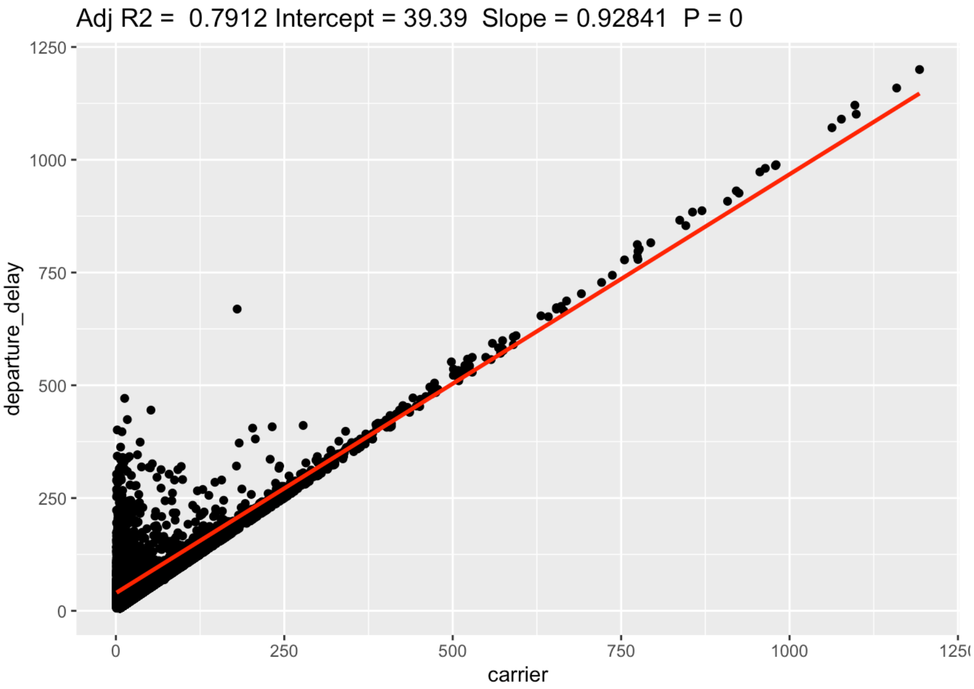


Figure 9. *The Adjusted R-squared Value for Carrier Delay is 0.7912*

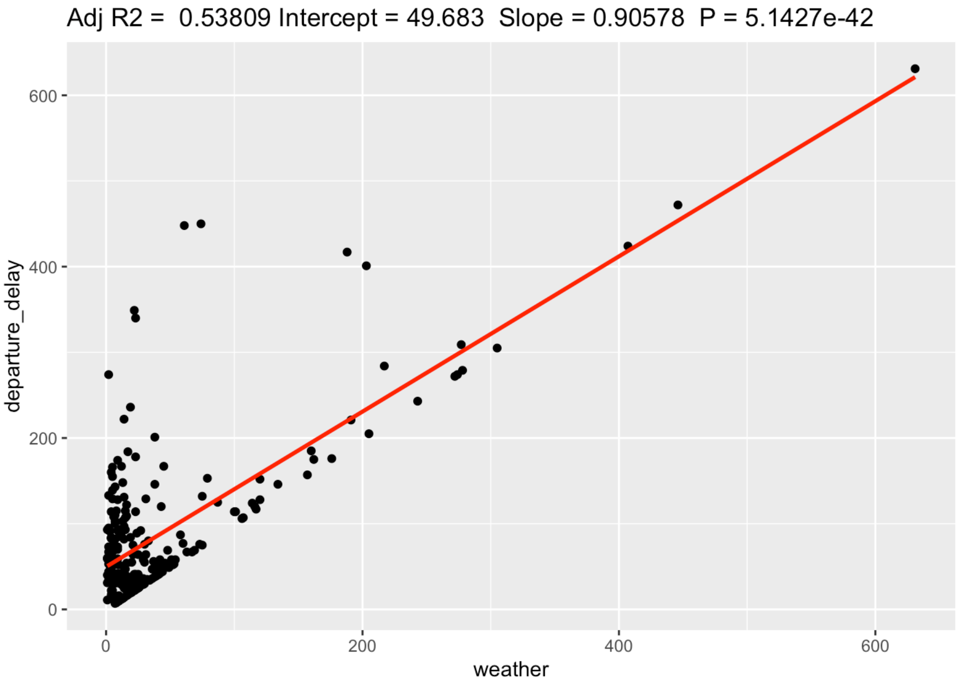


Figure 10. *The Adjusted R-squared Value for Weather Delay is 0.5381*

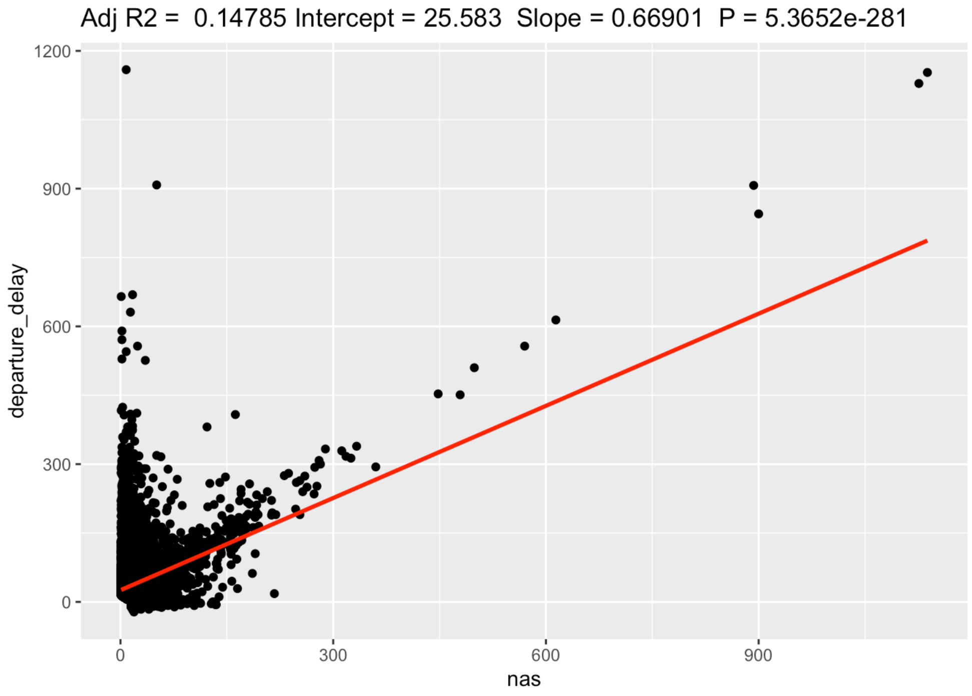


Figure 11. *The Adjusted R-squared Value for NAS Delay is 0.1479*

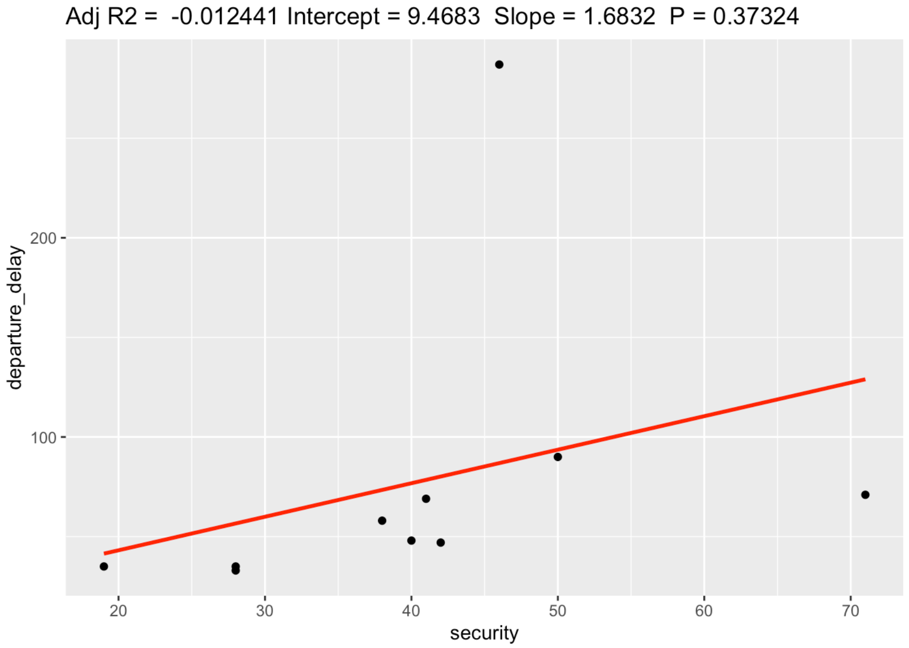


Figure 12. *The Adjusted R-squared Value for Security Delay is -0.0124*

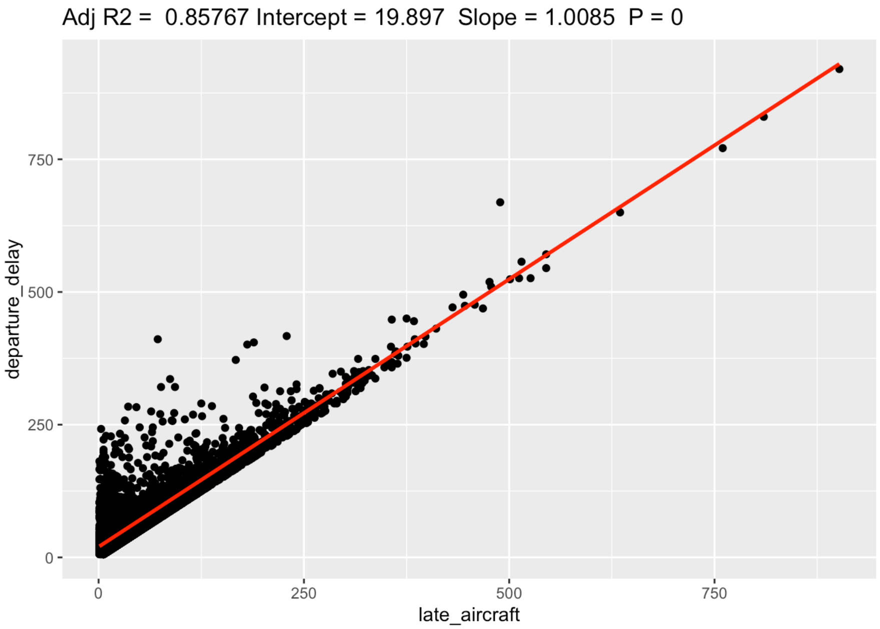


Figure 13. *The Adjusted R-squared Value for Late Aircraft Delay is 0.8577*

During my analysis, I came up with another hypothesis that using a dataset where there are only delays would have a better prediction. To do so, I used the original dataset and removed the rows that have all 0s, meaning flights that had no delay in all five predictors regardless of their departure time difference and here are the results.

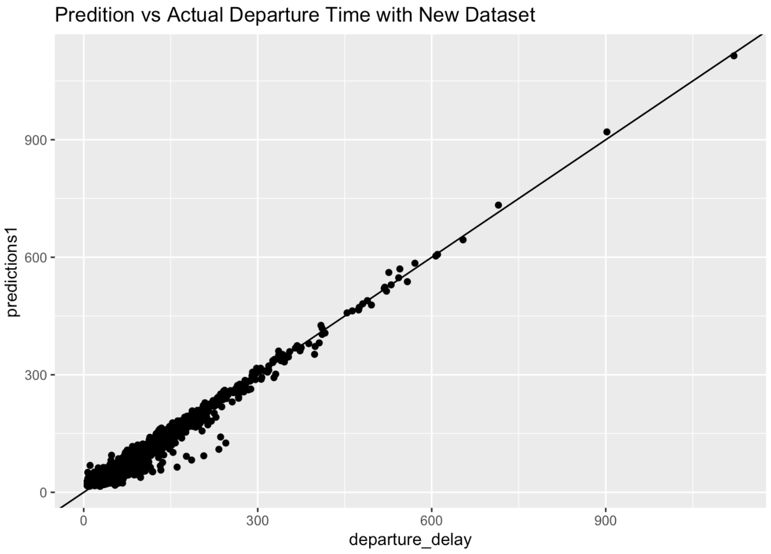
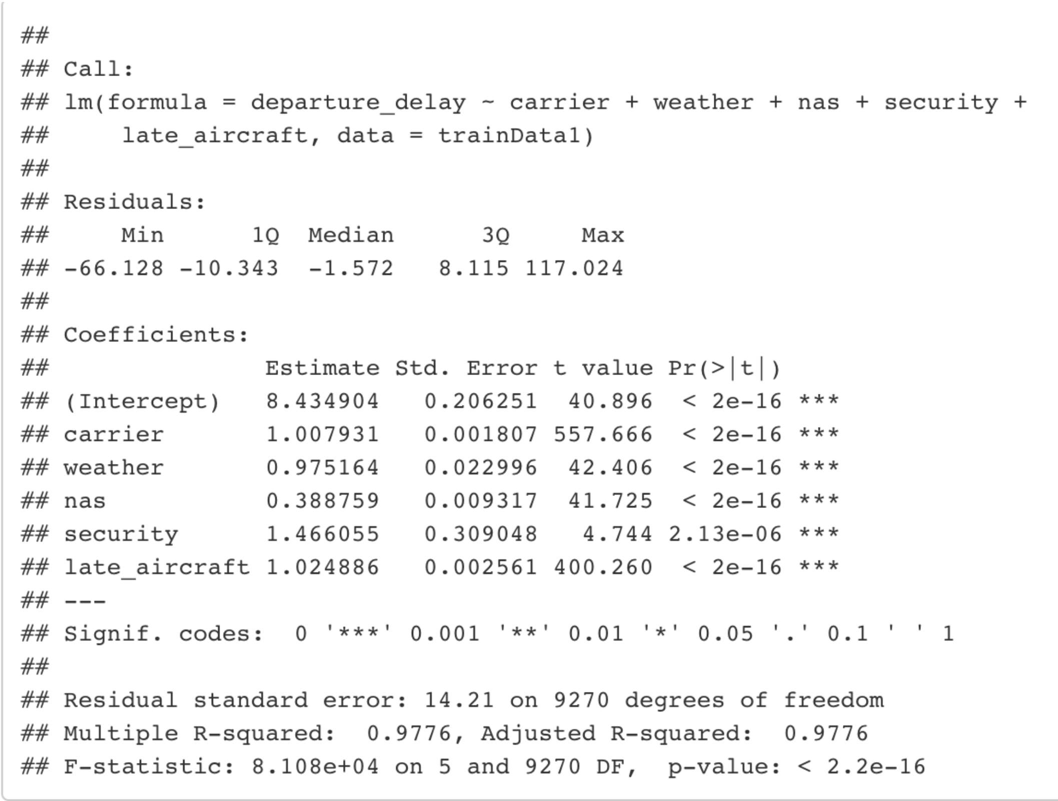


Figure 14. *Prediction Graph With New Dataset*



```
##
## Call:
## lm(formula = departure_delay ~ carrier + weather + nas + security +
##     late_aircraft, data = trainData1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.128 -10.343  -1.572   8.115 117.024
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.434904   0.206251  40.896  < 2e-16 ***
## carrier       1.007931   0.001807 557.666  < 2e-16 ***
## weather       0.975164   0.022996  42.406  < 2e-16 ***
## nas           0.388759   0.009317  41.725  < 2e-16 ***
## security      1.466055   0.309048   4.744 2.13e-06 ***
## late_aircraft 1.024886   0.002561 400.260  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.21 on 9270 degrees of freedom
## Multiple R-squared:  0.9776, Adjusted R-squared:  0.9776
## F-statistic: 8.108e+04 on 5 and 9270 DF,  p-value: < 2.2e-16
```

Figure 15. *Summary of Model with New Dataset*

I observe that the range of residuals has significantly reduced and the intercept has now become positive. Meaning that there is an average 8 minute delay whenever there is a delay in any of the predictors. It is also interesting because NAS and weather both have a coefficient of less than 1, this could be due to the fact of how weather delay is measured, because weather delay is when weather conditions at the departure airport, destination airport or during the flight. For example if there was turbulence, this would be considered a weather delay but turbulence only happens when the plane is flying, meaning it's after the plane has already departed. From looking at the T and P value, I can state that security, NAS and weather delay would not be considered an important factor and the reason being the T value is moderately low, as I explained before with the previous summaries. This also matches with my results when I plotted each predictor individually, the adjusted R-squared values for these three predictors were not as high as carrier and late aircraft delay.

Other than the reduction of range in residuals, another good news is that the Adjusted R-squared value is higher than the previous model at 0.9776, this implies that the model with the new dataset makes a better prediction.

4. Discussion/Conclusion

To sum it up, my model did a great job at predicting flight departure times when considering the major causes of flight delay, with an accuracy of 92.47%. Through mathematical analysis, I was able to conclude that for my data set of analyzing departure times, within 2013 to 2023 for flights in San Francisco international airport solely from Delta Airlines, the main factors of flight delays were due to late aircraft and carrier issues. This is an extremely valuable resource as it gives airlines a direction to work towards when improving their on-time performance. For example, these issues are well within the airline's control, if they could perhaps improve their scheduling or operational procedures, then perhaps maintenance or crew problems would be minimized. My model is also easy to understand and it's able to clearly depict these relationships between the predictor and response variables. However, this also determines the downfall of my model.

If my model was to consider data that were to include a non-linear relationship, I would not be able to capture these non-linear patterns. In particular, my model was considered fair because I only considered data from a specific airport, so factors such as weather delay would apply to all flights objectively. Nonetheless, if my model was to have a parameter that would take weather conditions into account, my model wouldn't perform well. This is because although it may be intuitive that less favorable weather conditions will result in a flight being delayed, this is not always the case. For instance, a shower of light rain will usually not result in a delayed flight, but a severe thunderstorm would result in a flight being heavily delayed, or even canceled. This implies that the opportunity to build a more accurate model would require a different approach outside the scope of my class, utilizing different regression methods. Furthermore, my model would be more accurate if I were to include more predictors.

For example, to include flights that ended up being canceled, or even factors that would indicate that a delayed flight could be out of their control.such as a late arriving passenger, damaged runway, etc.

5. References

United States Department of Transportation Bureau of Transportation Statistics. Detailed statistics departures. 2013-2023, Departure Delay, Cause of Delay, San Francisco, CA: San Francisco International (SFO), Delta Airlines Inc. (DL)<https://www.transtats.bts.gov/ontime/departures.aspx>

United States Department of Transportation Bureau of Transportation Statistics. Airlines operate more flights in june; on-time performance hits a high. 2020.

United States Department of Transportation Bureau of Transportation Statistics. Flight cancellations stabilize in May, but total flights hit another record low. 2020.

Anish M Kalliguddi and Aera K Leboulluec. Predictive modeling of aircraft flight delay. Universal Journal of Management, 5(10):485–491, 2017.

Evangelos Mitsokapas, Benjamin Sch ̈afer, Rosemary J Harris, and Christian Beck. Statistical characterization of airplane delays. Scientific Reports, 11(1):7855, 2021.

Ronald Wesonga, Fabian Nabugoomu, and Peter Jehopio. Parameterized framework for the analysis of probabilities of aircraft delay at an airport. Journal of Air Transport Management, 23:1–4, 2012.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Johnston, S. (2015, April 23). A quick and easy function to plot LM() results with GGPLOT2 in R. Johnston Lab. Retrieved May 2, 2023, <https://sejohnston.com/2012/08/09/a-quick-and-easy-function-to-plot-lm-results-in-r/>

Zhao, L. S23-MATH150-Lecture22, April 23, 2023, <https://catcourses.ucmerced.edu/courses/27033/files/folder/LectureSlides?preview=5948414>

Carreira-Perpinan, M.A.. CSE176 Introduction to Machine Learning — Lecture notes, September 2, 2019, <https://faculty.ucmerced.edu/mcarreira-perpinan/teaching/CSE176/lecturenotes.pdf >

Bhat, H. MATH 140 Lecture Note (Hand-Written), February, 2022

Shams, A. (2022, May). *Why do airlines always fly so early? why do I have to be at the ... - quora*. Why do airlines always fly so early? Retrieved May 1, 2023, <https://dailybest.quora.com/https-www-quora-com-Why-do-airlines-always-fly-so-early-Why-do-I-have-to-be-at-the-airport-at-4-am-I-honestly-dont-und>

6. Code for the Mathematical Model

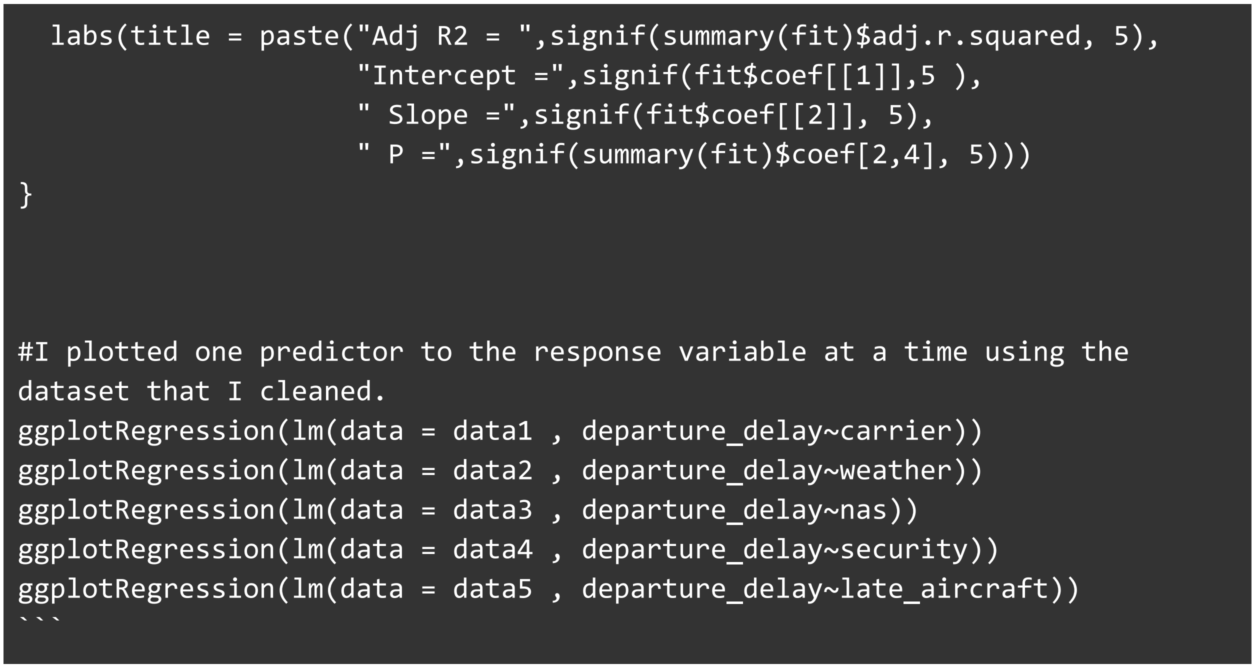
~~~
---
title: "MATH150Project"
author: "Shaun Tran"
date: "2023-04-26"
output:
  html_document: default
  pdf_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, warning=FALSE, message = FALSE)
~~~

```{r}
# Load necessary libraries
library(tidyverse)
library(caret)

# Read in data
data <- read_csv("Detailed_Statistics_Departures1.csv")

# Select relevant columns
data <- data %>%
  select(DelayCarrierMinutes, DelayWeatherMinutes,
DelayNationalAviationSystemMinutes, DelaySecurityMinutes,
DelayLateAircraftArrivalMinutes, DestinationAirport, DeparturedelayMinutes)

# Rename columns for convenience
colnames(data) <- c("carrier", "weather", "nas", "security",
"late_aircraft", "destination", "departure_delay")


```

```{r}
# Split data into training and test sets
set.seed(123)
trainIndex <- createDataPartition(data$departure_delay, p = 0.8, list =
FALSE)
```

~~~
trainData  <- data[trainIndex, ]
testData   <- data[-trainIndex, ]
```

```{r}
# Create linear model using least squares method on training data
model <- lm(departure_delay ~ carrier + weather + nas + security +
late_aircraft, data = trainData)

# Make predictions on test data
predictions <- predict(model, testData)

# Add predictions to test data
testData$predictions <- predictions

# Plot predictions against actual values using ggplot
ggplot(testData, aes(x = departure_delay, y = predictions)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1) + ggtitle("Prediction vs Actual
Departure Time")

summary(model)
```

```{r}
#model without security
model12 <- lm(departure_delay ~ carrier + weather + nas + late_aircraft,
data = trainData)
summary(model12)
```

```{r}
#Removing 0s from the predictor column because I am only looking at when
there is a delay in the factors.
data1 <- trainData[trainData$carrier != 0, ]
data2 <- trainData[trainData$weather != 0, ]
data3 <- trainData[trainData$nas != 0, ]
data4 <- trainData[trainData$security != 0, ]
data5 <- trainData[trainData$late_aircraft != 0, ]
```

```{r}
#data6 is the new dataset that contains only the rows with a delay in at
~~~

```
least one predictor
data6 <- data[rowSums(data[c(1, 5)] != 0) > 0,]
set.seed(123)
trainIndex1 <- createDataPartition(data6$departure_delay, p = 0.8, list =
FALSE)
trainData1 <- data6[trainIndex1, ]
testData1 <- data6[-trainIndex1, ]
modelgg <- lm(departure_delay ~ carrier + weather + nas + security +
late_aircraft, data = trainData1)
predictions1 <- predict(modelgg, testData1)
testData1$predictions1 <- predictions1
ggplot(testData1, aes(x = departure_delay, y = predictions1)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1) + ggtitle("Prediction vs Actual
Departure Time with New Dataset")


summary(modelgg)
```
~~~
```{r}
#Here i did a fitted model without the predictor  security
mod1 <- lm(departure_delay ~ carrier + weather + nas + late_aircraft, data
= trainData)
#anova is a function to analyze the variance, this is to see their
F-statistic and P value.
anova(mod1,model)
~~~
~~~
```{r}
#this function allow me to plot and show the adjusted r square, intercept,
slope and p value
ggplotRegression <- function (fit) {

require(ggplot2)

ggplot(fit$model, aes_string(x = names(fit$model)[2], y =
names(fit$model)[1])) +
  geom_point() +
  stat_smooth(se=FALSE, method = "lm", col = "red") +
~~~



```
  labs(title = paste("Adj R2 = ",signif(summary(fit)$adj.r.squared, 5),
                     "Intercept =",signif(fit$coef[[1]],5 ),
                     " Slope =",signif(fit$coef[[2]], 5),
                     " P =",signif(summary(fit)$coef[2,4], 5)))
}



#I plotted one predictor to the response variable at a time using the
dataset that I cleaned.
ggplotRegression(lm(data = data1 , departure_delay~carrier))
ggplotRegression(lm(data = data2 , departure_delay~weather))
ggplotRegression(lm(data = data3 , departure_delay~nas))
ggplotRegression(lm(data = data4 , departure_delay~security))
ggplotRegression(lm(data = data5 , departure_delay~late_aircraft))
```
```
```