36-401 DA Exam 1

Sarah Li (sarahli)

October 14, 2022

Introduction

Airline services constantly struggle to fulfill customer satisfaction pertaining to the accuracy of flight information. A common complaint made by passengers is the unexpected delay at the arrival destination. The industry wants to gain a better insight of factors contributing to delayed arrivals in order to better provide good services at optimized prices. Specifically, the Bureau of Transportation Statistics (BTS) has been tracking every flight's statistics for the last two decades, broken down into: flight times, taxiing, distance, departure or arrival delay, etc. **(1)** In this report, we will specifically analyze the relationship between flight arrival and departure delay and assess the linearity between the two variables. Additionally, we will also observe if their relationship is affected by the presence of weather conditions.

From our analyses, we are hesitant to conclude that there was evidence of a linear relationship between flight arrival delay and departure delay, even after data transformation. Additionally, weather doesn't seem to play a role in the relationship either.

Exploratory Data Analysis and Initial Modeling

For this report, we will be parsing in a sample of 4887 flight records from the "airlines.csv" file. We will individually assess the distributions and important statistics for Departure Delay (the main predictor variable) and Arrival Delay (the response variable). **(2)** From the histograms displayed in Figure 1, we can see the Departure Delay minutes are unimodal but heavily skewed right. The minutes range from -29 minutes (negative indicating early departure) to 1099 minutes (delay in departure). The center, or median, is at -1 minutes, indicating that over half the recorded flights departed earlier than expected. **(3)** As for Arrival Delays, the distribution of minutes is also unimodal but heavily skewed right. The minutes range from -60 minutes (negative indicating early arrival) to 1092 minutes (delay in arrival). The center is at -2 min, indicating that over half of the flights arrived early to their destinations.

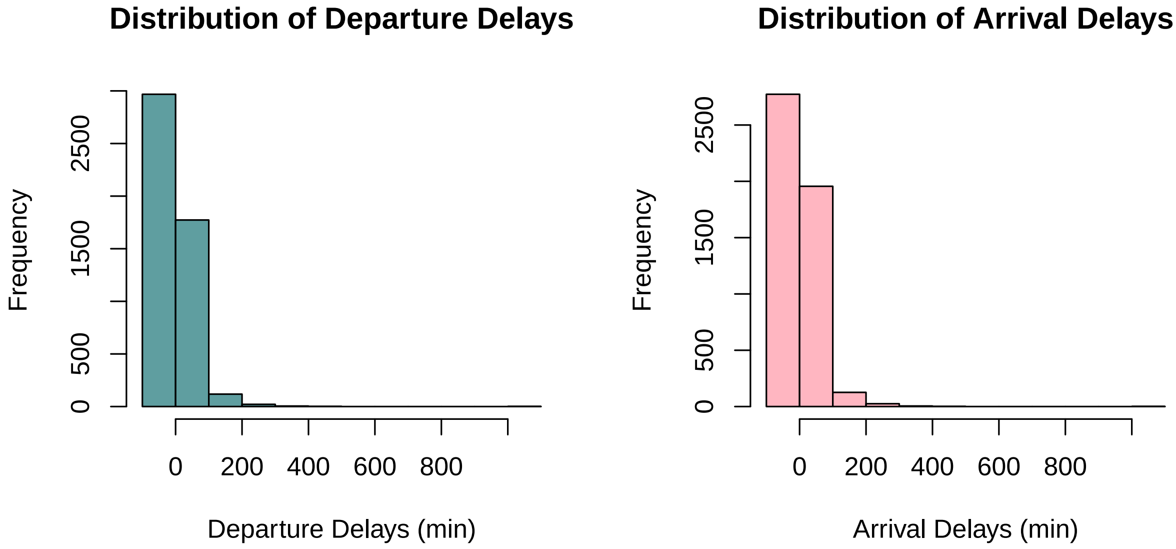

Figure 1: EDA Analysis of Departure and Arrival Delays

Due to heavy skew, we proceed to transform the data by stretching it outward. Subsequently, we log both Arrival Delay and Departure Delay, accounting for their minimum values (-60 and -29 minutes, respectively) by shifting them up by 61 and 30 minutes. This is to ensure that every data point is included, as the log function only returns finite values for inputs greater than 0. **(2)(3)** Shown in Figure 2 below, though the transformed distributions aren't perfectly symmetric, they are much more normalized and easy to work with.

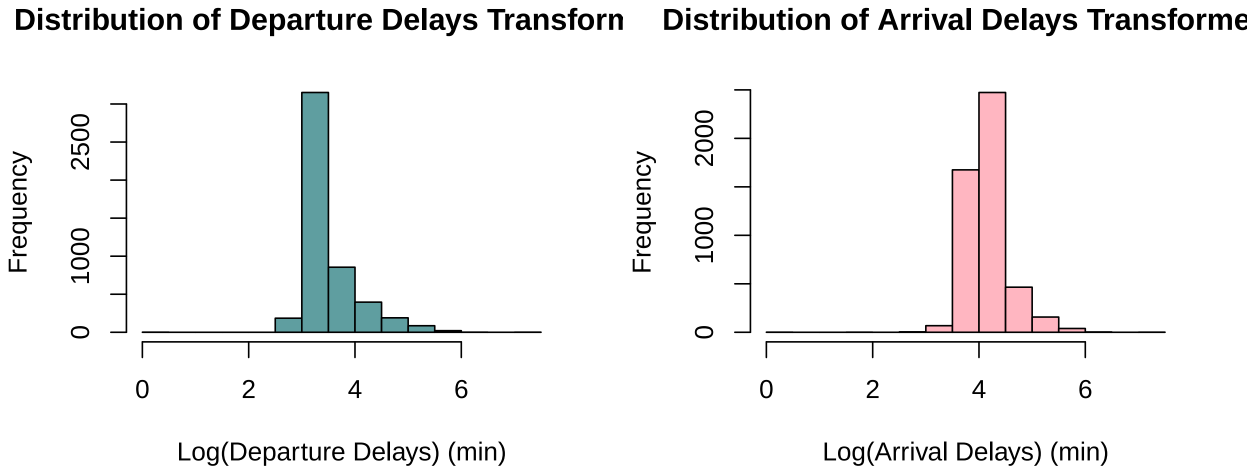

Figure 2: EDA Analysis of Departure and Arrival Delays (Transformed)

(4) Deciding to work with the log transformations of both Departure Delay and Arrival Delay, we will perform a bivariate EDA on the initial modeling of their relationship from the raw data and compare it with their logged counterparts. From the original scatterplot(left) in Figure 3, we can see there is an obvious outlier on the top right of the graph at (1092 minutes, 1099 minutes). We have decided to plot the relationship between delays and the estimated regression line with and without this outlier to see if there was high leverage (confirmed there wasn't and plot omitted).

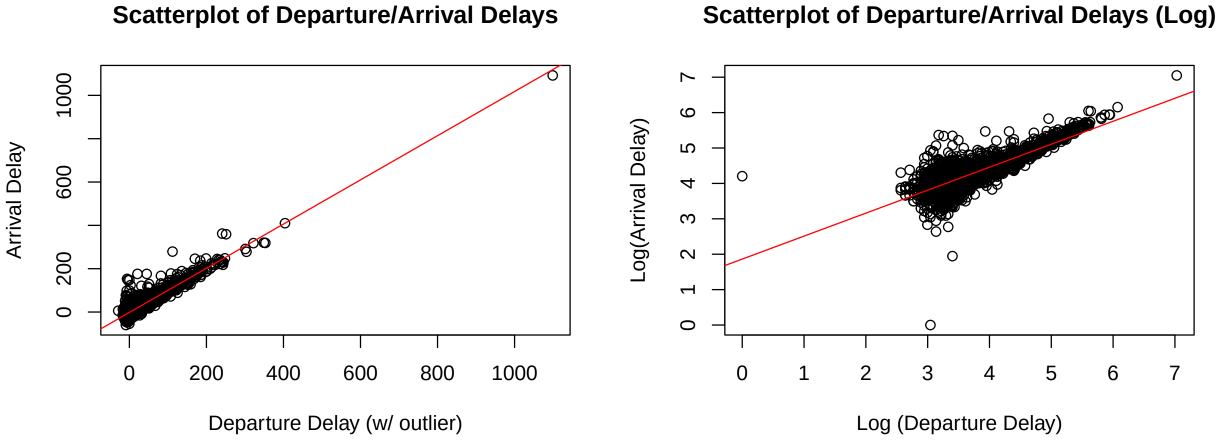

Figure 3: Scatterplot of Departure and Arrival Delays

Additionally, from Figure 3 above, it is visually evident that Log(Arrival Delay) and Log(Departure Delay) have a positive and moderately strong linear relationship. Both plots show that many of the points float above the general trend, potentially leveraging the model in the upward direction. The points also clutter around values below 100 minutes for both axes. However, there seems to be a few outliers with comparably lower values of Log(Arrival Delay) and Log(Departure Delay) to the left half of the scatterplot (~3) that extend the plot left. From the linear model summary of the log transformation, we observe a correlation coefficient of .8244, which indicates a relatively strong correlation between Log(Arrival Delay) and Log(Departure Delay) (though not necessarily linear).

Diagnostics

(5) The model we have chosen to construct our regression model is: **log(Arrival Delay + 61) ~ B0 + B1 x log(Departure Delay + 30)**. For our sample case, $\hat{B}_0 = 1.8628$ and $\hat{B}_1 = 0.6488$.

(6) One major assumption of our model is that the flight records are random and independent. Since we have taken a sample from two decades of records, and flight data are not associated, this assumption can hold. To handle our assumptions of linearity, we want take a closer look at outliers and points that could influence our analyses. From the calculation of influential points using Cook's Distance metric, we can see from Figure 4 that there is 1 main outlier that exceeded the median of the distribution, determined by comparing the values to the quantiles of the F distribution with 2 and n-1= 4885 degrees of freedom. of Log(Arrival Delay) and Log(Departure Delay) that could affect the regression model. The main point was indeed very far from the main cluster of points but doesn't seem to leverage the data too much.

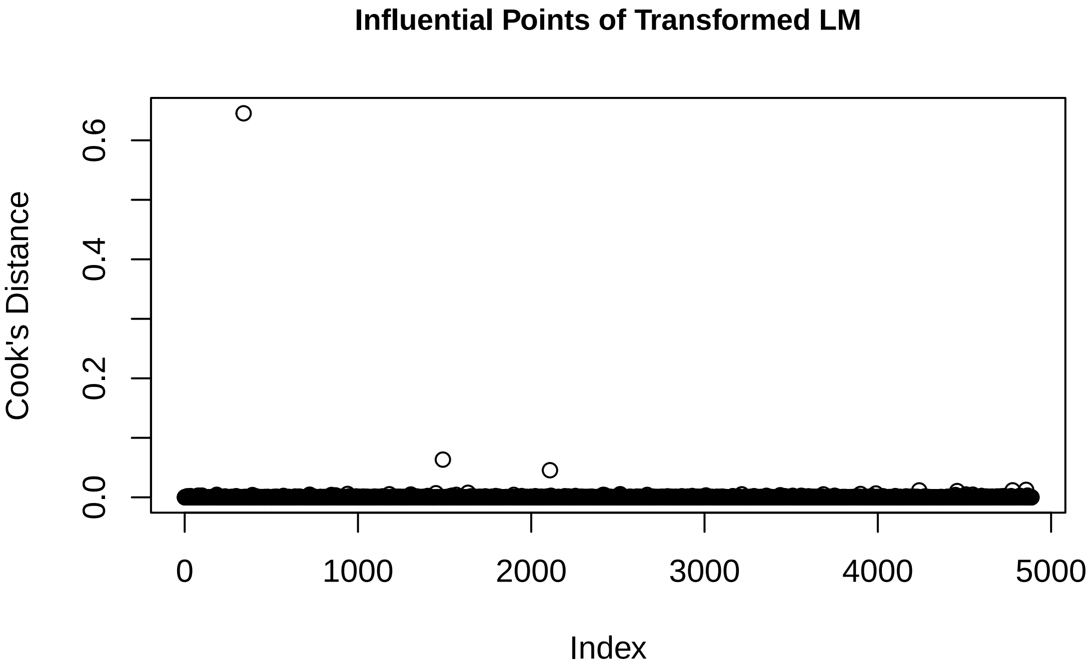

Figure 4: Cooks Distance for Transformation

(6 cont.) We will proceed by looking at the residuals (errors) of our model. From Figure 5, we can see that the residuals follow a similar pattern as the scatterplot. The points clutter around the middle of the plot and thin out toward the right; the main outlier, labeled 340, also leveraged the residual line upward (though it minorly imapacts the regression line). If our model assumptions are correct, there should be no relationship between the residuals and predictor. However, based on the figure, we can see that the residuals are not evenly distributed and does not follow a homoskedastic scatter about 0. Even if we were to run the residuals excluding the outliers taken from cook's distance, the clustering pattern cannot be ignored. Thus, the residuals violate out assumptions for a linear model.

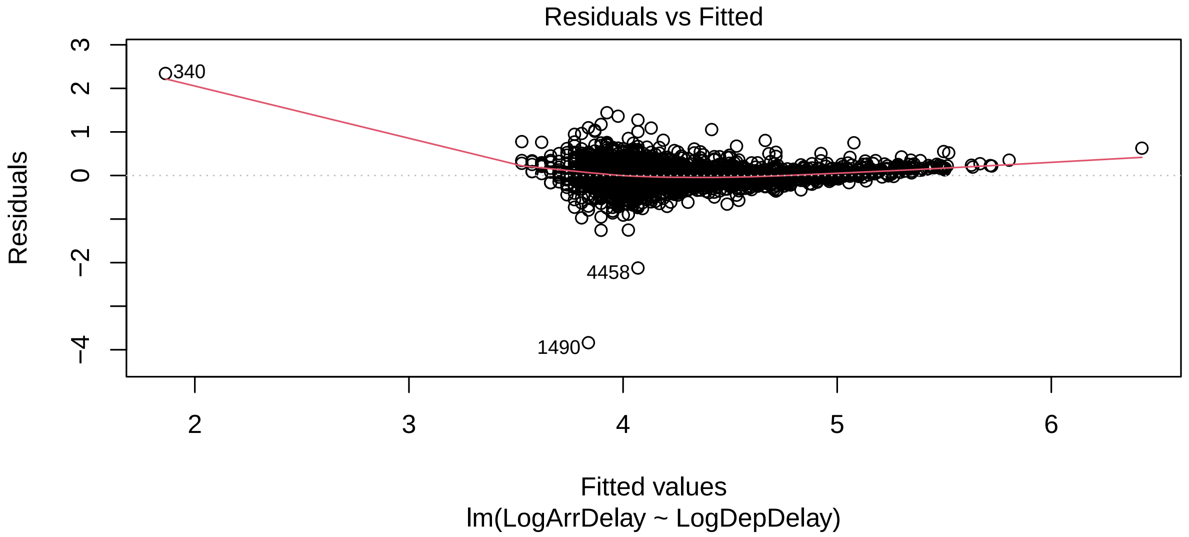

Figure 5: Residuals Plot of the Fitted Model

To back up this observation, we proceed to plot the Normal Q-Q plot of the fitted model with and without the outlier of concern at 340. We observe in Figure 6 that the points in both Q-Q plots do not closely follow the diagonal line near the ends but still have the same shape and pattern. In other words, removing the outlier did not change much.

Thus, after using Cook's Distance to isolate influential data points, and then confirming through our estimated regression line, residuals, and Normal Q-Q plot that the single outlier should not leverage our data in a sample size of 4887, we will keep the outlier in the data for further interpretation. **(7)** The assumptions that underlie our model are reasonable as demonstrated by the figures. There may not be a linear relationship between Arrival and Departure Delays but there is arguably an association factored in with other conditions.

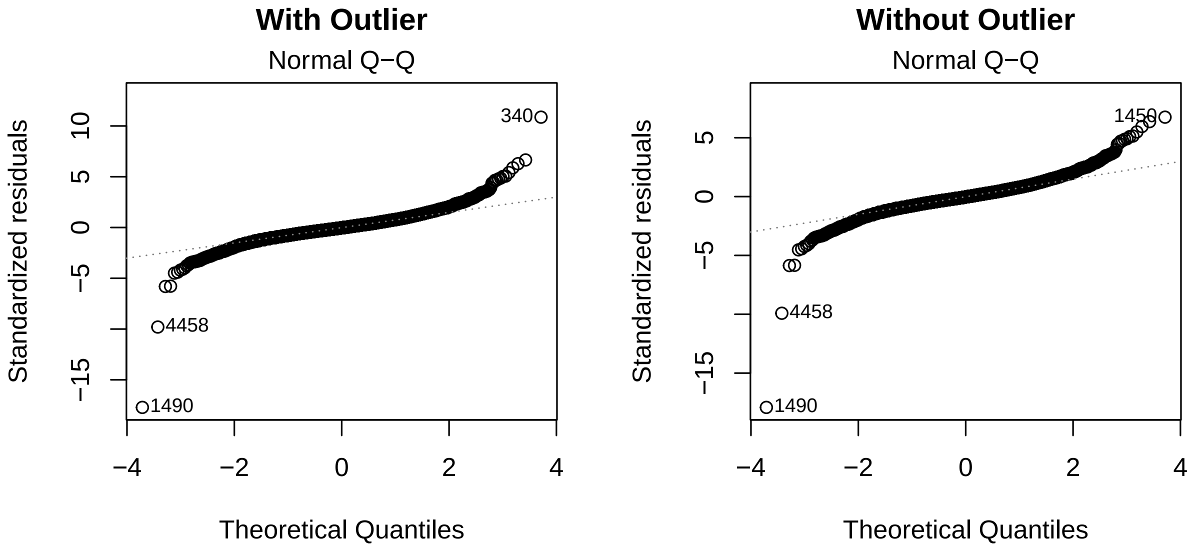

Figure 6: Normal Q-Q Plots of the Fitted Model

Model Inference and Results

We conclude that there is no simple linear relationship between Arrival Delays and Departure Delays for flights. That is, an change in unit of Departure Delay (minutes) can not necessarily predict Arrival Delay (minutes).

(8) In our regression model, however, there is strong statistical evidence of there being a relationship between the two variables. We set $H_0 : B_1 = 0$ and $H_0 : B_1 \neq 0$ as our null and alternative hypotheses for the simple linear regression model. H_0 tests for no association, a slope of 0, while H_A tests for an association. From our conducted hypothesis summary, we conclude a very small and statistically significant p-value < 0.05. We, in turn, reject H_0 and conclude that there is an association between the predictor (Departure Delay) and response variable (Arrival Delay) in our transformed log linear model.

(9) Given a mean Departure Delay of 200 minutes, we obtain an estimated arrival delay of 158.42 minutes. The 90% confidence for the expected value of Arrival Delay for all flights given a Departure Delay of 200 minutes is (153.962, 162.97). We are 90% confident that the true population mean of Arrival Delays for the population of Departure Delays of 200 minutes is between 158.42 minutes and 162.9701 minutes.

```
##        fit      lwr      upr
## 1 158.4199 153.9622 162.9701
```

(10) As stated in our introduction, we have to assess whether weather conditions affect the relationship between Departure and Arrival delays. We decided to study the flight delays when weather conditions were present and not present, separately.

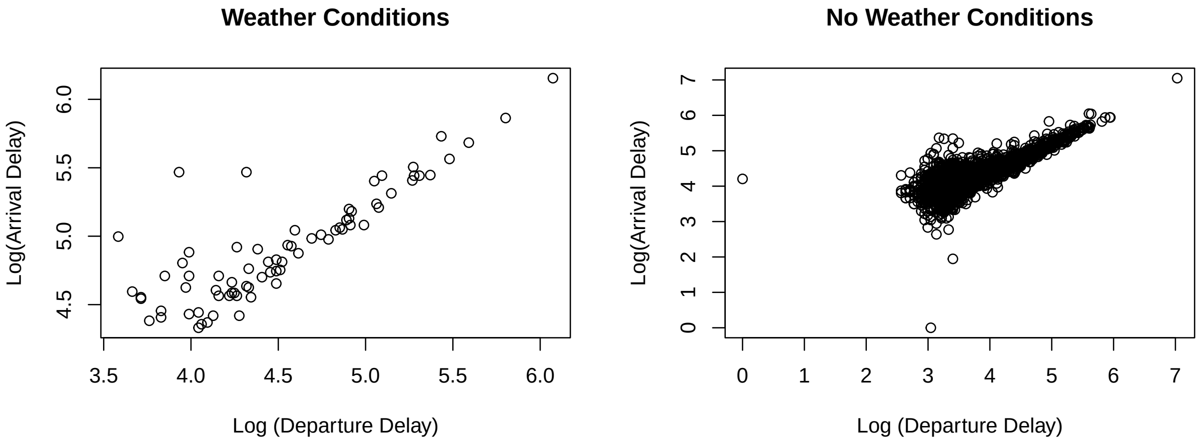

Figure 7: Transformed Delays w/ and w/o Weather Conditions

(10 cont.) In Figure 7, we observe that the relationship between Log(Arrival Delays) and Log(Departure Delays) both with and without weather conditions is positive and has a moderately strong linear relationship. After implicitly running the summaries of both models, we observe very small p-values (< 0.05 alpha), indicating a statistically significant relationship for our separately fitted models conditioning on weather disruptions. Subsetting on flights with weather conditions, we are 90% the true slope of the fitted model predicting Arrival Delay from Departure Delay is within (0.552, 0.704). For flights without weather conditions, we are 90% confident the true slope of the fitted model is within (0.63, 0.6516). However, since these confidence intervals overlap, we cannot conclude a significant difference in the fitted models between the relationship of Arrival and Departure delays given weather conditions. In other words, the presence of weather disruptions doesn't seem to affect a flight's relationship between Arrival and Departure delays for our model.

```
##                   5 %      95 %
## (Intercept) 1.7296093 2.4240025
## LogDepDelay 0.5520865 0.7043341
## Weather            NA        NA
```

```
##                   5 %      95 %
## (Intercept) 1.8507308 1.9275912
## LogDepDelay 0.6299204 0.6516073
## Weather            NA        NA
```

Conclusion and Discussion

(11) After conducting a thorough analysis of the relationship between Departure Delays and Arrival Delays for flights, we decisively conclude that there does not exist a linear relationship between the two variables. The reasoning follows from the violations of a simple linear regression. Even though we assumed our records were i.i.d and transformed both the predictor and response variables to have roughly normal distributions, the scatterplot and residuals vs fitted values demonstrate that the erroros are not normal and uncorrelated. There is evidence of heteroskedasticity about 0, with values fanning starting from low values of both Arrival and Departure Delays. Factoring the presence of weather conditions, we also cannot conclude a significant difference among flights with and without them. However, given relatively high correlation coefficients for th fitted transformed models, we can conduct further analyses and testing for the possibility of another type of association (i.e. Multiple linear regression)