36-401 DA Exam 1

Amor Ai (muai)

October 14, 2022

Introduction

Due to a high volume of complaints largely related to flight delays, the airline industry is struggling to keep their costs low while maintaining good relationships with their customers. In a effort to deliver improved service and keep their customers happy, we seek to help the industry understand and identify what features contribute to flight delays by answering the following questions: **(1)** Is there a relationship between the flight arrival delay and the departure delay? Does this relationship depend on if the delay was due to weather or not? Ultimately, we hope that answering these questions will shed light on the features associated with flight delays, in order for the industry to gain a better understanding of flight delays and hopefully use this information to help prevent them in the future.

Exploratory Data Analysis/Initial Modeling

Out of the multitude of variables that may be contributing the flight delays, we will specifically be focusing on 3 main features of flights tracked by the Bureau of Transportation Statistics in 2008: departure delay, arrival delay, and weather. The departure and arrival delay variables in our dataset are denoted in minutes, with negative values indicating early departures and arrivals. The variable weather describes either if there was a delay due to weather or if the delay was not weather related. Our sample includes 4887 flights in total. As a first step in our analysis in discovering a relationship between arrival delay and our main predictor departure delay, we must explore each variable individually. To do so, we use histograms to examine the distribution of our variables:

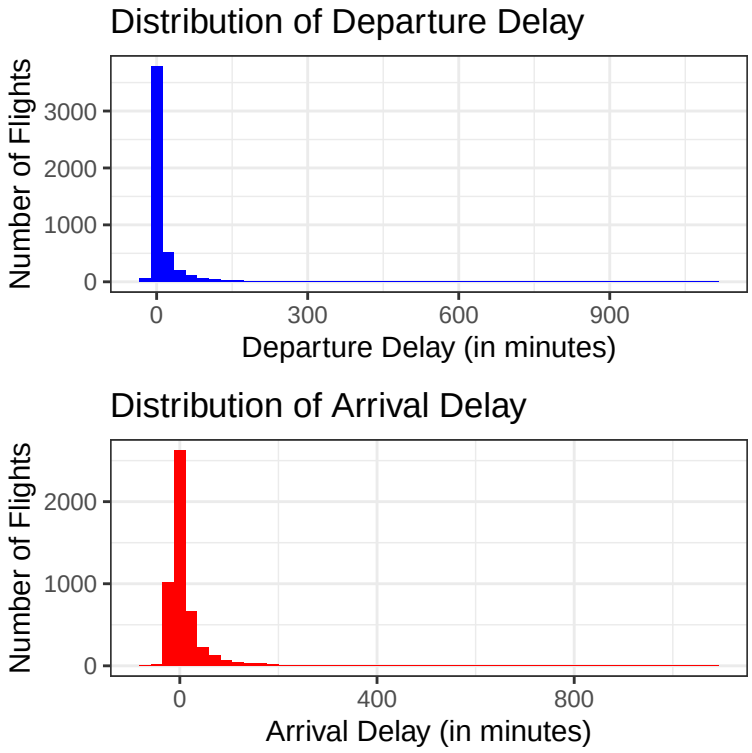

Figure 1: Histograms showing the univariate distributions of variables Departure Delay and Arrival Delay.

(2) The univariate distribution of the departure delay of flights is illustrated by the top plot in Figure 1; we see that the data is unimodal but severely skewed to the right, ranging from -29 minutes to 1099 minutes with an clear spike at around 0 minutes. **(3)** Similar to the departure delay, we see that the distribution of the flight arrival delay, ranging from -60 minutes to 1092 minutes, is also unimodal and right-skewed, with most of the data centered between -10 and 11 minutes. The skewness of these histograms may suggest that a transformation might be needed, which will be addressed when looking at model diagnostics later.

(4) Transitioning to bivariate exploration, we observe a fairly strong and positive linear relationship between departure and arrival delay, such that when departure delay increases in minutes, arrival delay increases as well (Figure 2). This scatterplot also gives us more insight into the possible outliers in the data, specifically highlighting a seemingly high leverage point at roughly 1000 minutes in both departure and arrival delay. In addition, there appears to be potential outliers on the left side of the plot where the arrival delay values are slightly higher than the general trend. These outlier points, especially the delay of approximately 1000 minutes, is evident in the univariate distributions of the variables as well, illustrated by the strong right skew.

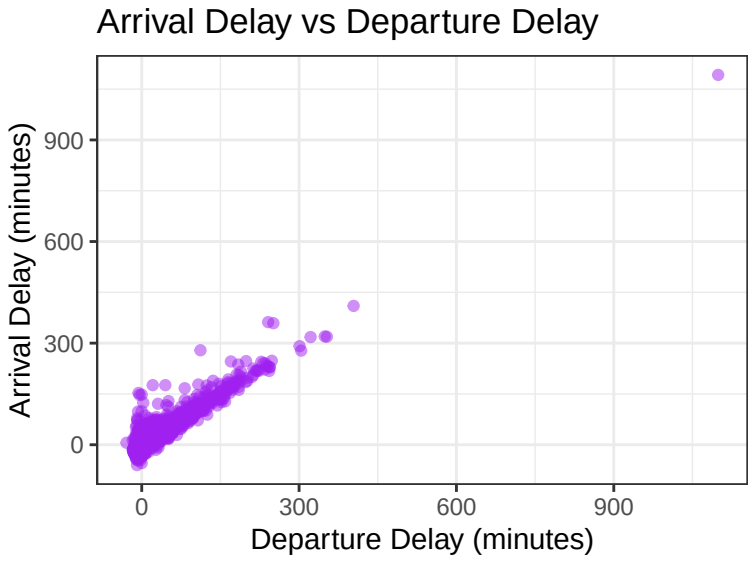

Figure 2: Scatterplot showing the relationship of flight arrival delay on departure delay

Diagnostics

After exploring and visualizing our variables of interest, there appears to be a linear relationship between flight arrival and departure delay — hence, the following linear model would be an appropriate starting point: Arrival Delay = β_0 + β_1*Departure Delay. To determine whether a linear model is actually the best fit for the data, we must first check model diagnostics by using residual analysis and checking for any influential points.

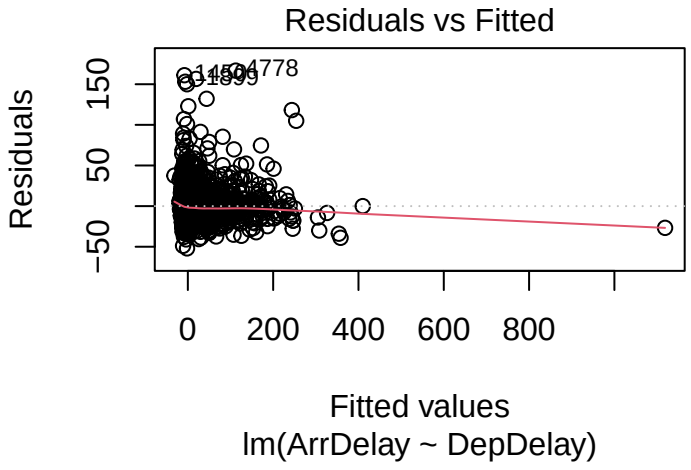

Figure 3: Diagnostics — Residual plot for a simple linear regression between flight arrival and departure delays

Looking at Figure 3, we observe that there are no obvious trends in the residuals, having relatively symmetric scatter about 0 (mean is approximately 0). Therefore, the linearity as-

sumption is reasonably justified — which is necessary for justifying the use of our model. However, when it comes to the spread of the residuals, there is evidently problems with heteroskedasticity as the residuals do not have constant variance. This problem seems to be affected by the one outlier with a notably larger fitted value on the right side of the plot. Consequently, we might want to check if this point is influential and fit the model again without it.

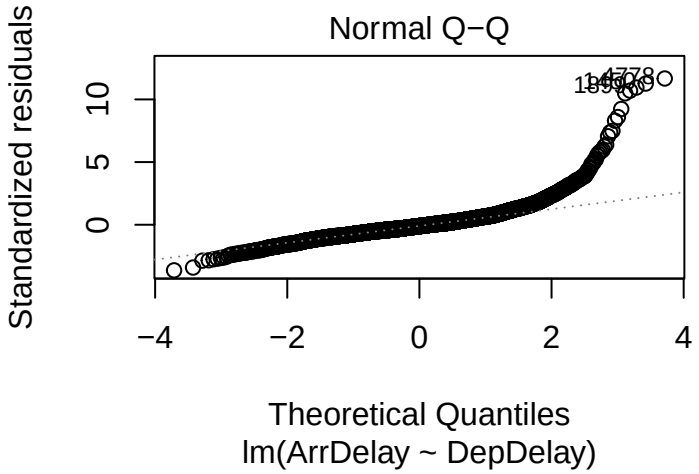

Figure 4: Diagnostics — Normal qqplot for a simple linear regression between flight arrival and departure delays

In the normal qqplot (Figure 4), we note some substantial deviation in the upper right side of the plot even though the rest of the points fall reasonably near the line on the qqplot. Again, this could be influenced by potential outliers in the data but is not as big of a concern since our sample size is large (n = 4887). However, these initial diagnostic results suggest that the residual assumptions are not all reasonably satisfied yet, and hence we have encountered the need to deal with outliers, try transformations, or even potentially choose a different model altogether.

As previously mentioned, there seemed to be an outlier that has the potential of influencing the fit of the model. We can formally check this by calculating the Cook's Distance, a measure that summarizes how much all the values in the regression model change when a potentially influencial observation is removed. Comparing the largest values for Cook's Distance to the quantiles of the F Distribution with 2 and 4885 degrees of freedom, we see that no observations exceed the median (50th percentile) of this distribution — perhaps suggesting that there is not a definite cause for concern. Nevertheless, the 2108th observation returned a percentile of 0.37 which is not only close to 0.5, but it is also considerably larger than all the other values of Cook's distance, as seen in Figure 5.

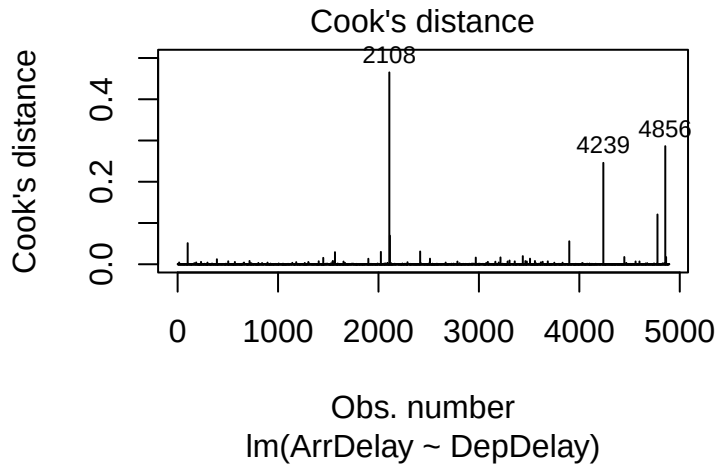

Figure 5: Plot of Cook's distance that estimates the influence of each observation

For further investigation, we can plot the residuals against leverages in another effort to identify influential points in our regression model. In a residuals vs. leverage plot, any points that fall outside of Cook's distance, indicated by dashed lines, is considered to be an influential observation. In Figure 6, since we can see that observation 2108 lies right on the dash line, we would be inclined to believe that this point is influential. Observations 4856 and 4239 are also relatively close to the dashed line because of their larger standardized residuals, however we would not identify these two observations as influential data points since their leverage is rather low and removing these observations from our data and fitting the regression model again does not change the coefficients of the model significantly.

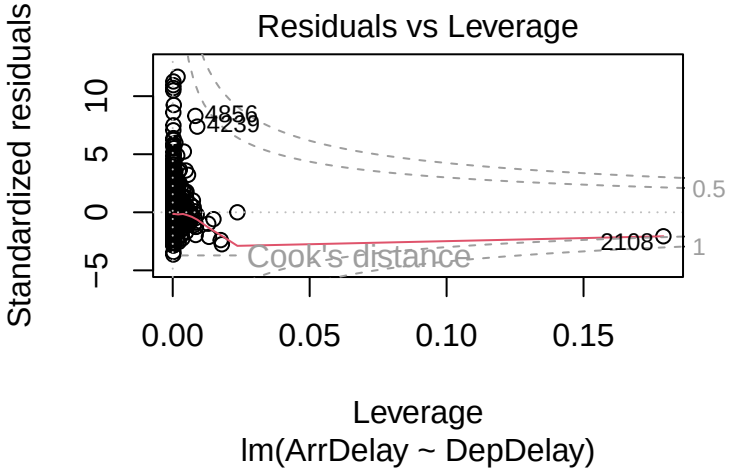

Figure 6: Plot of residuals against leverages to identify influential data points on the model

Before we move on, we should consider if our influencial observation 2108 is a result of a

mistake or not. Although the delays for that observation are far longer than the mean delay time, a departure and arrival delay of 1099 and 1092 minutes (~18 hours) respectively is plausible and might just be a rare occurrence. We can thus believe that the observation was not a result of a data entry error. To handle this outlier, we could simply remove it from the data; however before doing so, we should try to use a transformation of the predictor and response to try to "pull in" the outliers. Keeping our EDA in mind, this is one way we could address the the skewness of our data. When we transform only one of the two variables at a time (taking a log after adding a constant to shift the data since there are negative values), the relationship becomes nonlinear and thus should be not conducted. When we take the log of the departure and arrival delay (after shifting), the linearity holds, and it also slightly reduces the skewness of the distributions. However, these transformations not only make the interpretation of the model much more difficult, but it also does not substantially improve our model diagnostics. Evidence of this can be seen in Figure 7 as the residuals plot seems to violate both the linearity and constant variance assumptions. Therefore, in an effort to keep our model simple, we will chose not to use transformations but rather remove the overly influential observation instead.

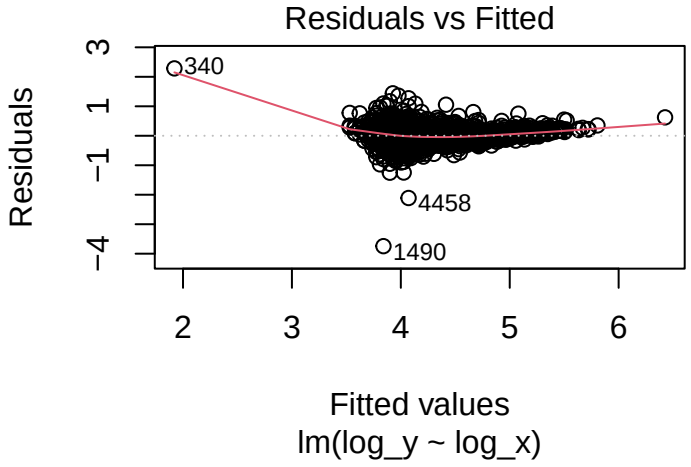

Figure 7: Residual diagnostic plot for a simple linear regression between log of flight arrival and log departure delays (with a shift)

(5) This is our final, chosen model for the relationship between flight arrival delay and departure delay:

$$ArrivalDelay = \beta_0 + \beta_1 * DepartureDelay$$

The linear relationship can be visualized in Figure 8. **(6)** This chosen simple linear regression model has the following assumptions: linearity (expectation of residuals = 0), constant variance (variance of residuals = σ^2), independence of errors (covariance of residu-

als = 0), and normality of errors (residuals are approximately normally distributed). The residual diagnostic plots of our final model are shown below:

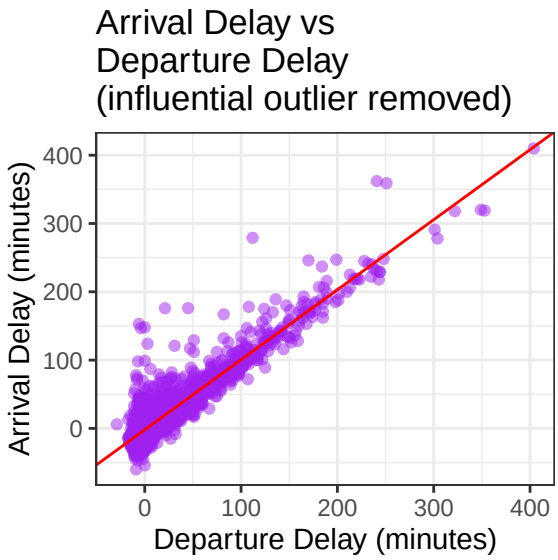

Figure 8: Scatterplot illustrating linear relationship between flight arrival and departure delay after removing the influential outlier

(7) As previously mentioned, the linearity assumption is reasonably justified by inspecting our scatterplot and seeing that the residuals do not present any obvious pattern.

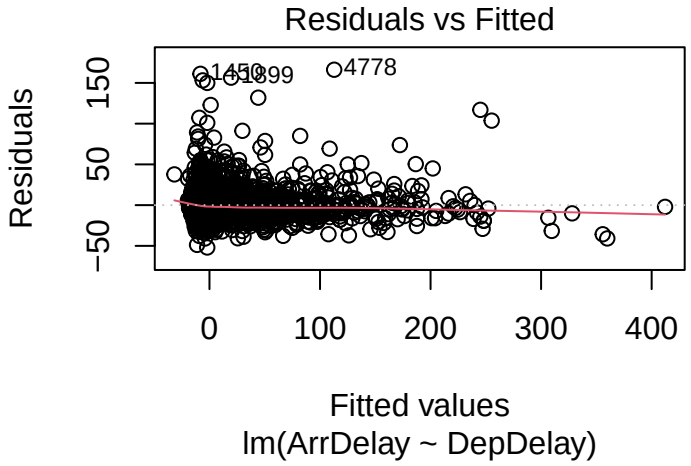

Figure 9: Residual plot for final SLR model between flight arrival and departure delays (influential outlier removed)

However, the residual plot seems to suffer from heteroscedasticity and the error distribution from the normal qqplot looks skewed by the presence of several outliers (heavy-tailed).

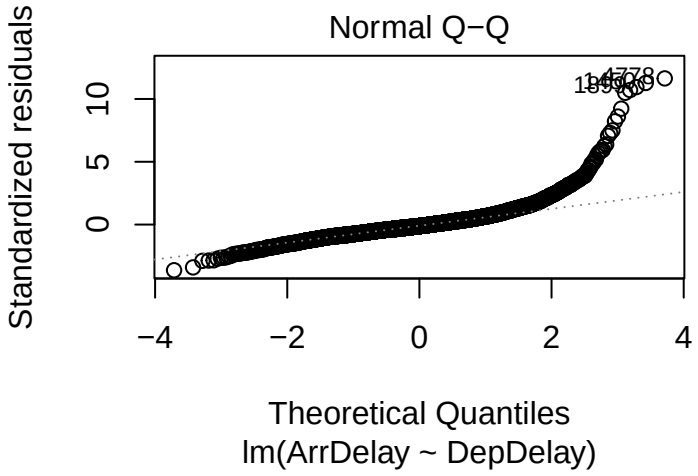

Figure 10: Normal qqplot for final SLR between flight arrival and departure delays (influential outlier removed)

Nevertheless, since the linearity assumption is the priority and addresses the overall fit of the model, we can still say that our model is appropriate since there are no signs of obvious nonlinearity between our variables. On the other hand, because of the considerable non-constant variance and non-normality of our data, our future inferences may not be valid and all confidence and prediction intervals must be interpreted with these assumption violations in mind.

Model Inference and Results

(8) To answer our main question of interest, we found, using our linear regression model, that there exists a statistically significant relationship between arrival and departure delay. Using a hypothesis test for the slope between the 2 variables, we know that the null hypothesis is that the slope between the variables is 0 (no association) and the alternative hypothesis is that the slope is not 0. Therefore, given an alpha level of 0.05, we reject the null hypothesis because the p-value (<2e-16) is so small. In context, we have sufficient evidence that the slope between flight arrival and departure delays is not 0, such that there is an association between the two. However, since there is substantial evidence of heteroskedastic and non-normal residuals (from our diagnostics above), the result and conclusion of this hypothesis test along with the estimated coefficients should be viewed skeptically.

(9) Using our model, we can predict that the estimated mean arrival delay for a flight which has a departure delay of 200 minutes is 203.02 minutes. A 90% confidence interval

for the expected value of the arrival delay for all flights which have a departure delay of 200 minutes is between 201.07 and 204.96 minutes. In spite of these results, it should be noted that violations of homoscedasticity and normality make it difficult for us to trust these values of the confidence interval.

(10) To answer the follow up question of whether this relationship depends on weather problems, we investigate the flights where there was a weather delay and flights where there was not. We find that there is a significant difference in our fitted models of the relationship between arrival delay and departure delay through comparing the confidence intervals of the parameters in the two models (differentiated by weather).

Table 1: 90% confidence intervals of the intercepts between arrival delay and departure delay between flights with and without weather delay

	Lower.Bound	Upper.Bound
Weather Delay	7.819421	24.191652
No Weather Delay	-2.475332	-1.789588

Table 2: 90% confidence intervals of the slopes between arrival delay and departure delay between flights with and without weather delay

	Lower.Bound	Upper.Bound
Weather Delay	0.8477218	1.002541
No Weather Delay	1.0112771	1.032344

As shown in the tables, both the estimated intercept and slope coefficient confidence intervals do not overlap between the weather delay and no weather delay groups. While this may suggest a significant difference, we again need to realize that there were violations in our model diagnostics and thus are unable to deem this finding completely valid. Nonetheless, it does provide some indication that the relationship between arrival and departure delays may indeed depend on weather.

Conclusion and Discussion

Airline complaints are a huge threat to the credibility of the businesses within the industry, as well as a burden financially. Since a majority of the consumer complaints are due to flight delays, we hope the airline industry can use our analysis to better understand flight delays to tackle the widespread problem. Using a sample of 4887 flights from Bureau of Transportation Statistics, we specifically wanted to discover any relationship between the flight arrival delay and the departure delay. Additionally, we wanted to explore if this relationship depends on weather problems. **(11)** From our analysis, we found that departure and arrival delay have a relatively strong, positive linear relationship, such that late departures are associated with later arrivals. While indeterminate, we also found that weather problems do seem to play a role in the delays — when the delay was due to weather, the estimated minutes of delay were larger than when the delay was not due to weather. However, like all inferences in our analysis, we should be skeptical of their validity since there were violations in our model assumptions. These violations indicate that the results of our analysis cannot be extrapolated and generalized, and that no formal conclusions or inferences can be confidently made. Despite this limitation, we can still use our chosen regression model to generate predictions within the range of our sample. Future steps could include fitting a different and more advanced model on the data in order to make valid statistical inferences or introducing additional variables in our model to help better predict and learn about arrival flight delays.