Data Analysis Report - Airline Arrival Delay

Kaili Chen

2018.10.12

- ## Introduction

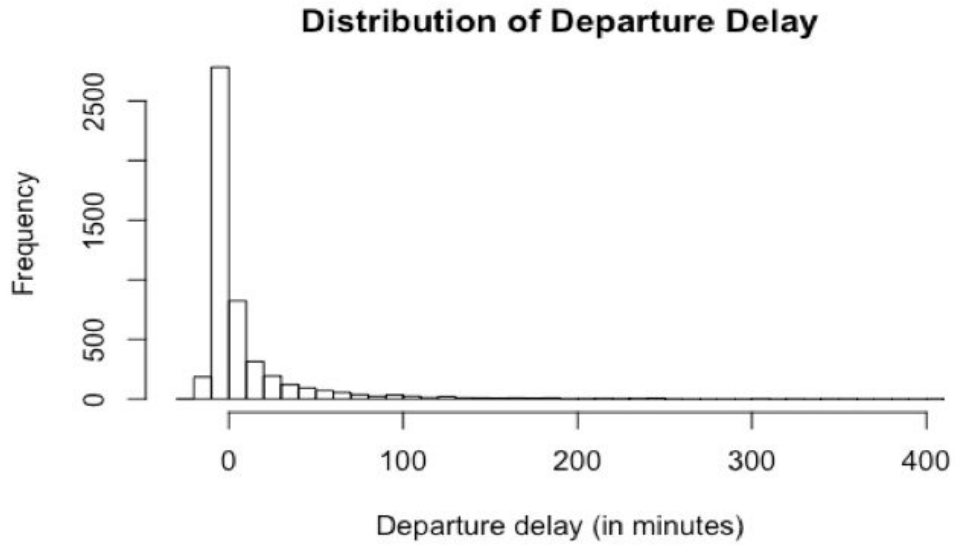
Q1 To help airline industry better understanding what factors contribute to flight delays, this data analysis report analyzes whether there is a relationship between the flight departure delay and the flight arrival delay. In addition, whether this relationship is dependent on weather problems is studied.

- ## Exploratory Data Analysis

The data used in this analysis includes 4887 flights sampled from 7,009,728 flights tracked by the BTS in 2008.

First, I did an EDA on the main predictor (departure delay) and the response variable (arrival delay). To clarify, the units for the predictor and the response are minutes, and negative values of arrival/departure delay means an early arrival/departure.

Q2 The predictor variable, Airline departure delay, ranges from -29 minutes (arrive 29 minutes earlier than scheduled time) to 1099 minutes (18.31 hours), with a mean of 10.21 minutes and a medium of -1 minutes.

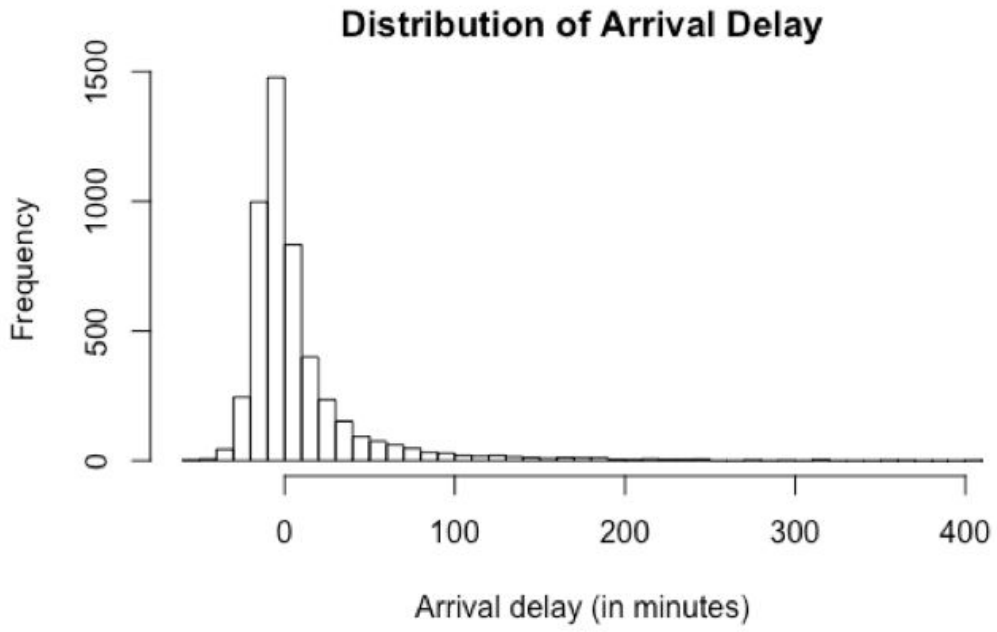


Q2

Based on the boxplot for the departure delay variable, we can tell that the maximum point 1099 is considered an obvious outlier possibly due to accidents happened to flight so it is removed from the data. From the R output figure above (Distribution of Departure Delay), the distribution of departure delay (after removing the outlier) is heavily skewed to the right, with a mode at -10 to 0 minutes. Therefore, I tried to do a log transformation on it. Since Departure delay ranges from -29 to 404, I took the log of (Departure delay + 50). Unfortunately, this transformation does not reduce the skewness significantly. Therefore, the original departure delay data after taking out the outlier is used later in the model.

Q3

Then I did the same EDA on Airline arrival delay data. It ranges from -60 minutes(-1 hour) to 410 minutes (6.83 hours), with a mean of 8.2 minutes and a medium of -2 minutes. From the R histogram output, the distribution of arrival delay is skewed to the right, with high concentration of data between -10 to 0 minutes.

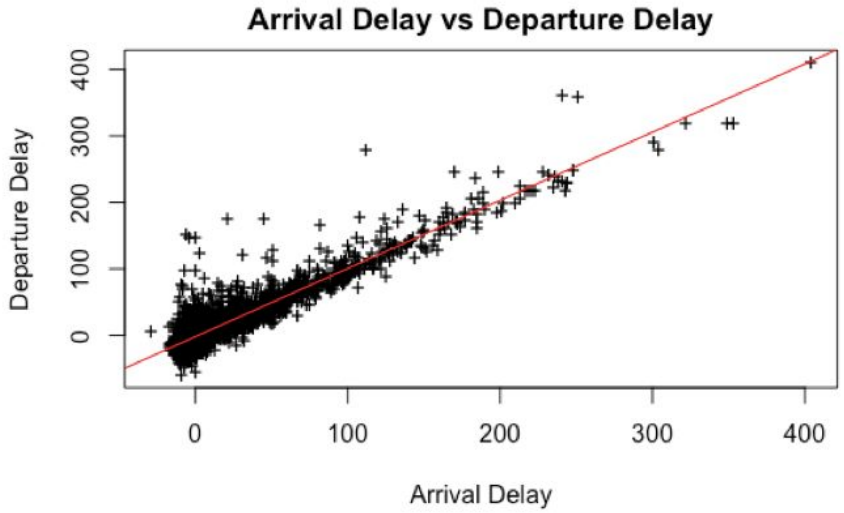


The scatter plot of arrival delay vs departure delay suggests a linear relationship between these two variables. Therefore, I fit a normal simple linear model on these two variables.

And then I took a look at the summary of the model in R. Below is a table for coefficients.

Table of coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.006045	0.213511	-9.396	<2e-16
DepDelay	1.025109	0.006131	167.209	<2e-16



After considering the fitness and complexity of the model as well as the randomness of the residual plot, I chose the following model as a final model to represent this relationship between airline arrival delay (Y) and airline departure delay (X):

Q4 Y = 1.025109 * X - 2.006045 + e

From R's output for the summary of this model, we can see that R-squared = 0.85 is a legitimate value to support a good model.

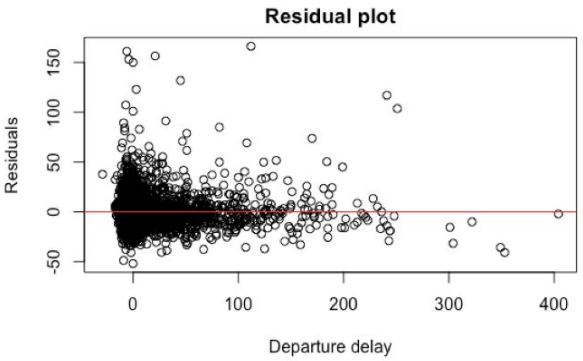
After I finalized my simple linear regression model, I went back to check the assumptions.The assumptions under this normal simple linear regression model are as follows:

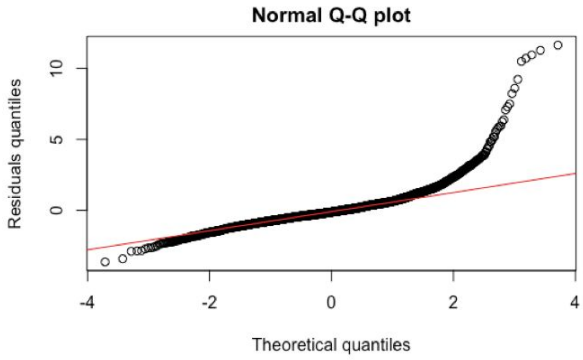
Q5

1) The distribution of departure delay is arbitrary.

2) If X(arrival delay) = x, Y(departure delay) = β1 * DepDelay(in minutes) + β0 + ε, for some constants β1 and β0 and some random noise variable ε.
3) E[ε|X=x] = 0 (no matter what x is) and Var(ε|X=x) = σ^2(no matter what x is)
4) ε is uncorrelated across observations
5) εi are iid normal

- **Diagnostics**





In order to check the assumptions and further diagnose potential problems with the fit of my regression model, a residual analysis is conducted here. When looking at the residual plot of my chosen model (Figure *Residual plot*), the residuals are generally equally scattered above and below 0. There is no obvious trend in the plot but it appears that the irreducible errors have a nonconstant variance across different X -- the variance becomes smaller as departure delay increase. Moreover, the normal Q-Q plot below indicates that the assumption of iid normal residuals may not be true. Therefore, the assumptions underlie my final model is not perfectly met.

Q6

After checking the assumptions, I ran a hypothesis test on β1 (the slope of the regression line).

Q7

H0: β1 = 0 (There is no relationship between these two models).
H1: β1 != 0

According to R output, p value is less than 2*10^(-16). Hence, there is sufficient evidence that there is a relationship between airline arrival delay and airline departure delay.

- ## Model Inference and Results

Q7 Based on the model constructed above, the estimated arrival delay for a flight which has a departure delay of 200 minutes is 203 minutes. R output displays the 90% confidence interval for the expected value of the arrival delay for all flights -- 201 minutes to 205 minutes.

Q8 We are 90% confident that for all flights that have departure delay of 200 minutes, the arrival delay is likely to be somewhere around 201 minutes (3.35 hours) to 205 minutes (3.41 hours).

As mentioned in the introduction, we are also interested in the effect of weather problems on this relationship between airline arrival delay and airline departure delay. Therefore, I constructed two new models using the same data, based on the weather variable (weather = 0 as not weather-related flight delay, or weather = 1 as flight delay due to weather).

Weather = 0: 95% CI(\hat{β1}) = (1.009258, 1.034364)

Weather = 1: 95% CI(\hat{β1}) = (0.8325564, 1.0177036)

Q9 Since there is a significant overlap between these two confidence interval(1.009 to 1.017) of \hat{β1}, we can conclude that there is no significant difference of the relationship arrival delay and departure delay between my fitted models.

Finally, from the chosen model, we can conclude that there is a linear relationship between airline arrival delay and airline departure delay. With one minute more in airline departure delay for all flights, the airline arrival delay is likely to increase by about 1 minute. Moreover, this relationship

Q10 between airline arrival delay and airline departure delay is not dependent on whether this flight delay is due to weather problems.