Stat 202, Project 2: Airlines

Michelle Hsieh
mhsieh2
There should be no collaborators

Due 4/18/18 at 11:59 pm

Contents

Introduction

I chose this research scenario because I'm interested in knowing what factors significantly affect plane arrival delay. It is believed that some days of the week have more delays than others and determining whether the day of the week is a significant factor would help people plan out their flights better. The overall modeling goal is to predict flight arrival delay time. We are trying to predict this to see whether the predictor variable significantly affects the arrival delay time. I would also like to see if there are other factors unaccounted for that may also affect the arrival delay time. The population of interest would be airplanes that have arrival delay times.

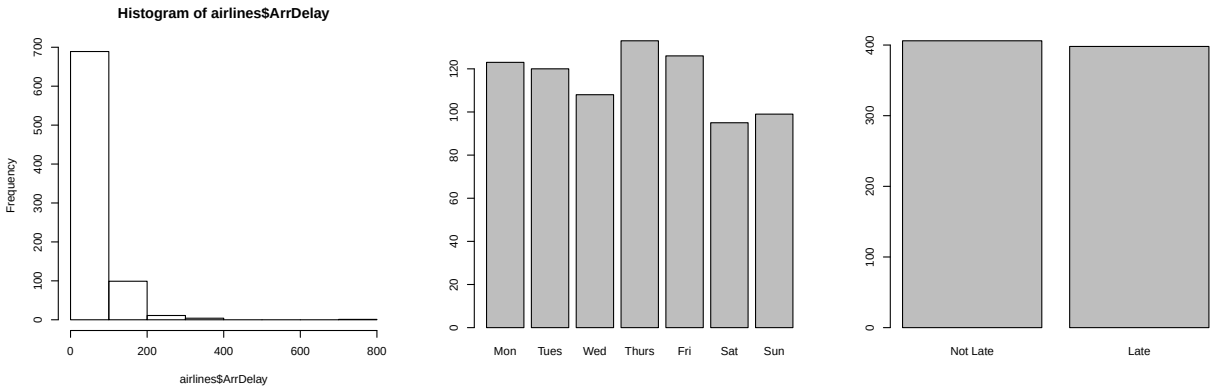
Exploratory Data Analysis

Data

ArrDelay	DayOfWeek	LateAircraft
31	Fri	Late
66	Wed	Late
26	Thurs	Late
37	Thurs	Late
18	Thurs	Not Late
25	Mon	Late

In the airlines data set, there are 804 observations with 3 variables. The variables are ArrDelay(arrival delay in minutes), DayOfWeek(day of the week), and LateAircraft(was there a delay? 1=Yes, 0=No). ArrDelay is the response. The predictors of interest are DayOfWeek and LateAircraft. We are interested in how much the predictors actually affect the arrival delay time.

Univariate exploration



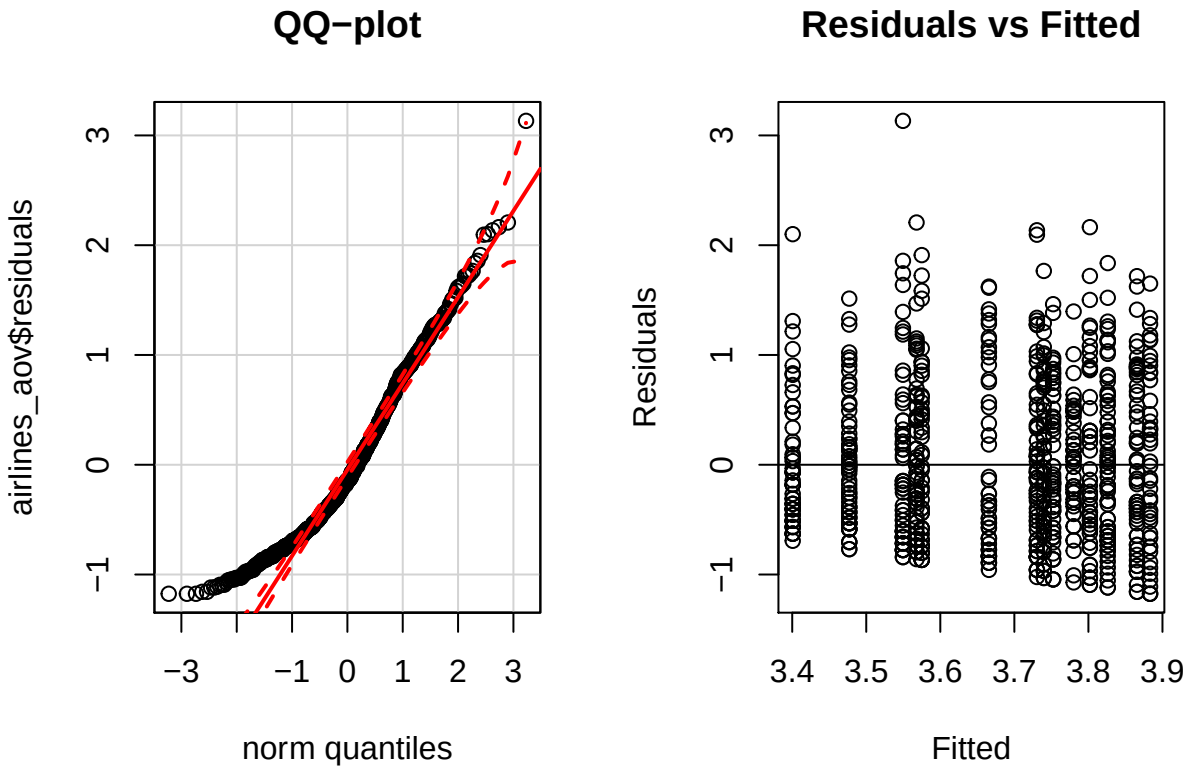
	Min	Q1	Median	Mean	Q3	Max	SD
ArrDelay	15	22	34	54.00871	66	798	55.77257

DayOfWeek	n
Mon	123
Tues	120
Wed	108
Thurs	133
Fri	126
Sat	95
Sun	99

LateAircraft	n
Not Late	406
Late	398

The graph for ArrDelay is skewed right with a mean of 54.0087 minutes and a standard deviation of 55.7725 minutes. I need to make further investigation to see if a transformation is needed. The graph for DayOfWeek seems pretty evenly distributed between all the days.Monday has a count of 123, Tuesday has 120, Wednesday has 108, Thursday has 133, Friday has 126, Saturday has 95, and Sunday has 99. The graph for LateAircraft is evenly distributed between the two categorical variables. Not Late has a count of 406 and late has a count of 398.

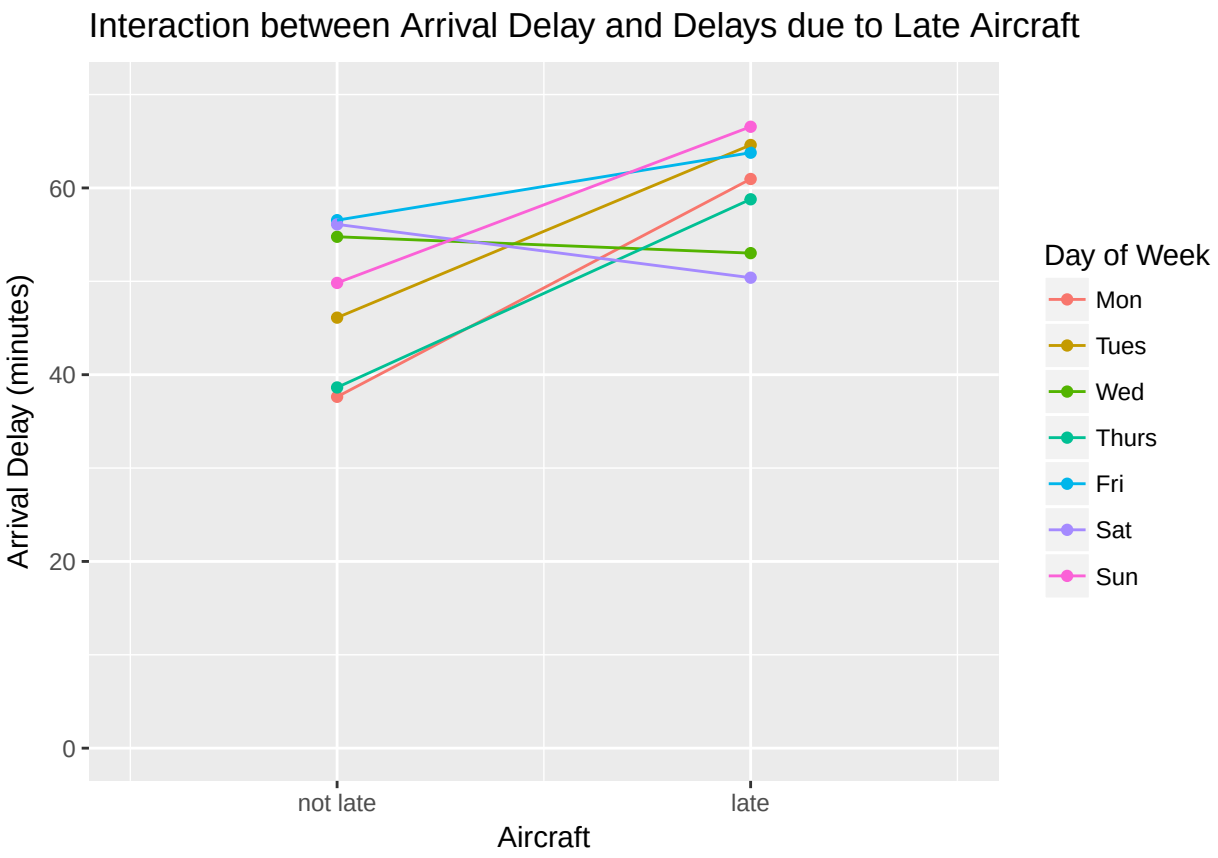
Assumptions



Based on the QQ plot, the data do not look normal because the data points go outside of the red bounded lines, therefore we cannot assume normality. We need to perform a transformation to make the data more normal. The log transformation made the points become more within the red line bounds. There are some points at the ends of the graph that do not fall within the red lines, but they can be overlooked since they're at the endpoints. Based on the residual plot,the residuals are roughly equally spread above and below the 0 line, so we meet the assumption that the errors have constant spread. The mean is also at 0 for the residual plot. Since the residuals are randomly scattered with no pattern, this shows that the data is independent.

Interactions and group means

Interaction plot:



In the interaction plot, almost all the arrival delay times increase when the previous aircraft is late. This shows that the late/not late instances could signficantly affect the arrival delay time. There are some intersections meaning that there is an interaction between arrival delay and late aircraft.

Table of means:

Table 5: Table of Means by Day (continued below)

LateAircraft	Mon	Tues	Wed
Not Late	3.40070071485336	3.56788186940179	3.73977844081159
Late	3.82613419756854	3.80195994968611	3.75236006733896
All	3.61	3.68	3.75

Thurs	Fri	Sat	Sun	all
3.47702678763224	3.54952933027606	3.66562277883968	3.57493017168472	3.57
3.73046444784346	3.86558296706971	3.78005358154938	3.88324580612988	3.81
3.6	3.71	3.72	3.73	3.69

The mean of the log of the arrival time delay for non-late aircrafts is 3.57 minutes and the mean of the log of the arrival time delay for late aircrafts is 3.81 minutes. Overall, the average mean of the log of arrival time delay is around 3.69 minutes.

Table of Standard Deviations:

Table 7: Table of SD by Day (continued below)

LateAircraft	Mon	Tues	Wed
Not Late	0.599126427318438	0.663065405880625	0.712432633782205
Late	0.731991399956498	0.811724864571422	0.657484635729641
All	0.09	0.11	0.04

Table 8: Table continues below

Thurs	Fri	Sat	Sun
0.563469145972458	0.806824186710395	0.827182056046161	0.772893122270807
0.761801050662893	0.749709635259326	0.524021717821644	0.800911993291917
0.14	0.04	0.21	0.02

all
0.1
0.1
0

Since the largest standard deviation/smallest standard deviation is less than 2 (1.6), we do not have to use the Keppel Correction.

Modeling

Table 10: Two-Way ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DayOfWeek	6	1.898	0.3163	0.6117	0.7211
LateAircraft	1	11.95	11.95	23.1	1.84e-06
DayOfWeek:LateAircraft	6	3.109	0.5182	1.002	0.4227
Residuals	790	408.5	0.5171	NA	NA

The S-pooled is the 0.5171ˆ0.5=0.7190967. The S-pooled is all the standard deviations of the DayOfWeek, LateAircraft, DayOfWeek:LateAircraft combined. Based on the two-way ANOVA table, only LateAircraft is a significant variable since it's the only variable that has a p-value of less than 0.05 (F-value=23.1, p-value=1.84e-06). DayOfWeek is not significant because its p-value is greater than 0.05 (F-value=0.6117, o-value=0.7211) and the interaction effect is not significant(F-value=1.002, p-value=0.4227). Since the interaction effect is not signficant, I will remove the interaction variable from the model and test the other two main effect variables.

Table 11: Two-Way ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DayOfWeek	6	1.898	0.3163	0.6117	0.7211
LateAircraft	1	11.95	11.95	23.1	1.838e-06

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	796	411.7	0.5172	NA	NA

In this model, DayOfWeek is still not a significant main effect (F-value=0.6117, p-value=0.7211). LateAircraft is a significant main effect (F-value=23.1, p-value= 1.838e-06). Since DayOfWeek is not signficant, we will remove it and test a One-Way ANOVA model using LateAircraft.

Main effects plot

Table 12: One Way ANOVA Model

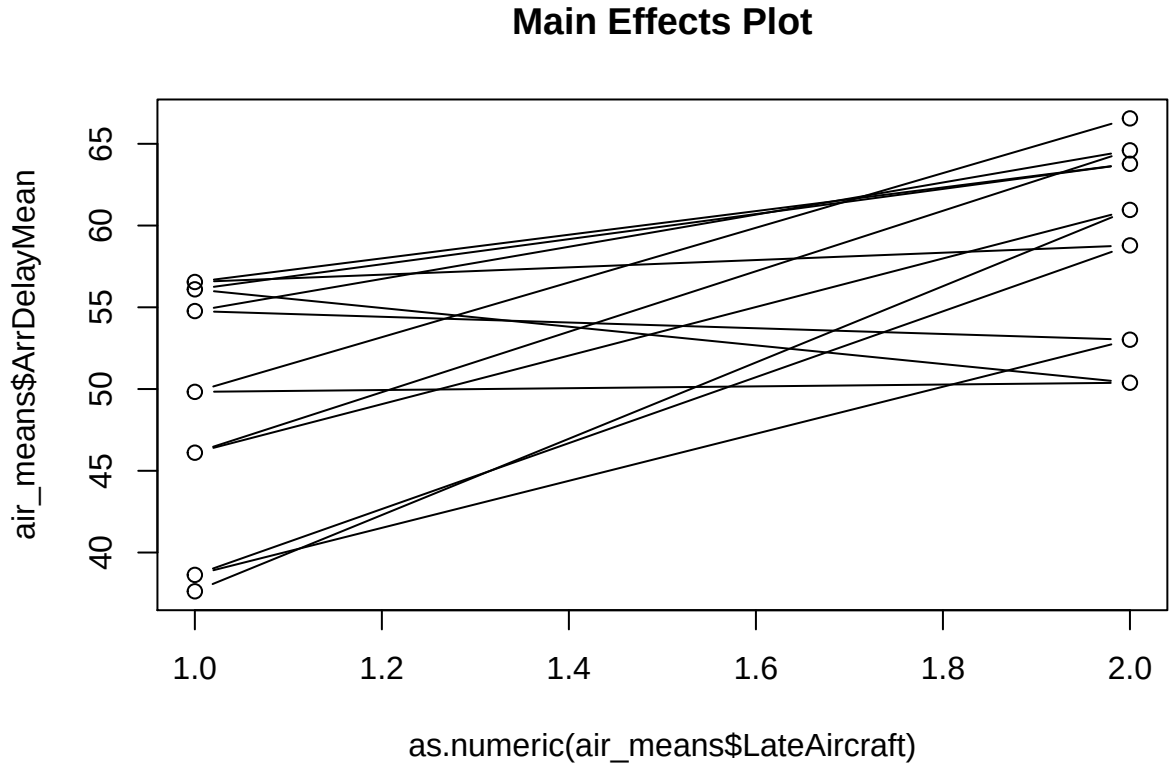
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
LateAircraft	1	11.76	11.76	22.79	2.15e-06
Residuals	802	413.7	0.5159	NA	NA

Leaving the interaction effect and DayOfWeek out of the model, LateAircraft remains a signficant main effect (F-value=22.79, p-value=2.15e-06). The final model would be:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

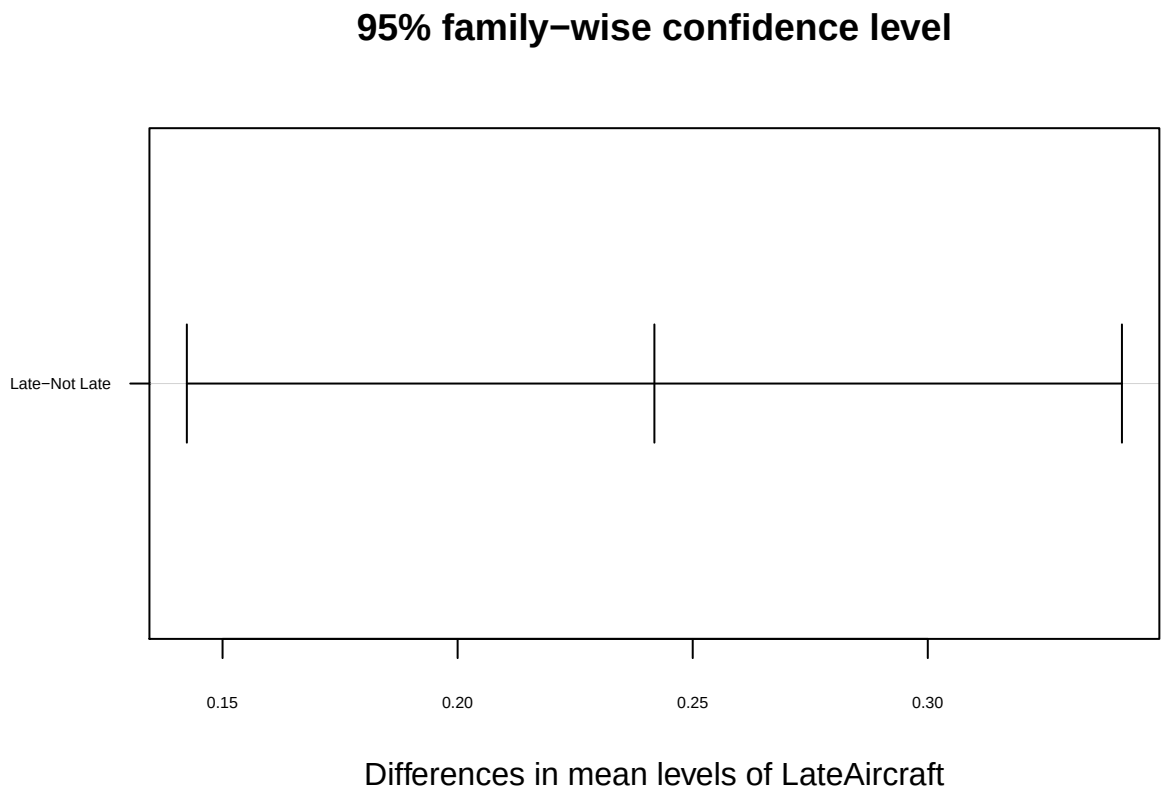
j=1,2(LateAircraft, Yes, No), i=flights There are three parameters in the model: Late, Not Late, and standard deviation.

Main Effects Plot



The main effects plot shows that there is a signficant interaction between the log arrival delay in minutes and the late aircraft variable because of the many intersections in the plot. In general, there seems to be an increase in arrival delay when an aircraft is late.

For each interval:



The question that we want to answer with the LateAircraft confidence interval is: Is there a difference in mean arrival delay time when the previous aircraft is late versus when the previous aircraft arrives on time? The confidence interval is:(0.1405214, 0.3431806). Calculation: (3.804-3.5627)+-2x0.7182x((1/398)+1/406)ˆ0.5 Since the interval does not contain zero, we conclude that the late aircraft main effect is significant. I am 95% confident that the true mean difference between late vs. arrivalDelay and not late vs. arrivalDelay is between the interval (0.1405214, 0.3431806). I can conclude there is a difference in arrival delay time when an aircraft is late versus when an aircraft is not late.

Discussion

In this section, describe your overall results (5-6 sentences).

- What was significant? What wasn't? What did you learn with respect to the original research questions (go back to the original scenario)? Did anything surprise you? What other information would you have liked to know? Please stick to discussions relevant to your data.

The lateness of the previous aircraft main effect is significant. The nature of the interaction between arrival delay time and lateness is positive since the arrival delay time would increase if the previous aircraft was late. The day of week for the flight main effect and the interaction effect were not signficant because the p-values (p-value of Day of Week= 0.7211, p-value of interaction effect=0.4227) were both greater than 0.05. I learned that there are many variables that could affect arrival delay time since the residuals are still very

large, this means there are other factors that affect arrival delay time. I was surprised that the day of week was not significant because I thought that busier flight days such as the weekends would have more delays than weekdays would. I would have liked to know more about what factors would affect the arrival delay time such as the weather condition, technical problems with the machinery, and congestion in air traffic. We could predict the arrival delay time more accurately if we could analyze the data for these additional factors.