STP 429 - Regression Analysis

Arizona State University

Takahiro Wada

Lab #1

Executive Summary

The airline industry is massive in scale with each airline rigorously competing with many other airlines. Airlines rely on arriving on time which allows efficiency and maximizing their profits hence it is important to understand what causes arrival delays. This allows airlines to manage their flights more efficiently and have more flights taken within a certain period of time. Key factors for arrival delays which have multiple factors range from carrier delays to weather delays. If delays are high then less flights are taken having an adverse effect on profits for airlines. To maintain consistent flights an airline must hedge delays by managing factors they can control or manage their flights better.

Many factors determine arrival delays for airlines, so a study was conducted to try and predict major delay factors in airlines. Factors that are considered critical when predicting delays are carrier delay, weather delay, National Air System delay, security delay, and late aircraft delays. A statistical analysis was performed to determine which delay factors are most important when building a model to predict arrival delays.

Using regression techniques to help analyze each of their factors and relevance, I can conclude that only carrier delay, weather delay, and late aircraft delay can be used to predict arrival delays. The following report will include details of the analysis. The model that was developed will help airlines manage how these delay factors impact their business.

Introduction

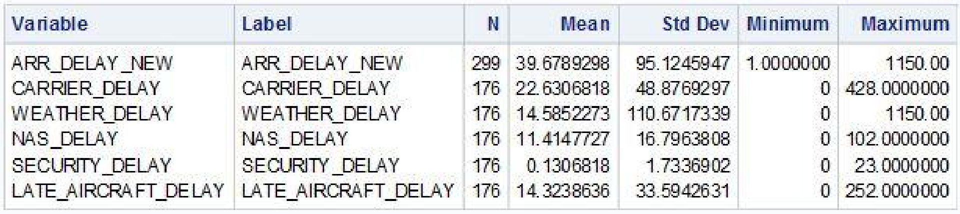
This study was generated in order to determine if there are factors that affect the arrival delay times so we can build a prediction model for airlines. We have five variables that may have a strong correlation to arrival delays such as carrier delay, weather delay, NAS delay, security delay, and late aircraft delay which are included for our prediction model in our study. Airlines rely on arriving on time in order to maximize their profits within a certain period of time.

Analysis

_______To determine the best model for predicting arrival delays, I first had to find if there is a strong association between the independent variables vs. arrival delay. The correlation matrix allows us to determine which independent variables have a strong association to our dependent variable. The chart allows us to see the strength for each variable also to see if we can identify any outliers within our data set. Once we decide which independent variables are acceptable to be used, I ran a regression analysis to find our best prediction model for determining arrival delay in airlines with our independent variables. Finally, I used a stepwise method, so the analysis will stop when all independent variables have been confirmed and are acceptable to be used for the model.

Data Section

From Table 7, the descriptive analysis for our 6 variables used for our model:



Variable	Label	N	Mean	Std Dev	Minimum	Maximum
ARR_DELAY_NEW	ARR_DELAY_NEW	299	39.6789298	95.1245947	1.0000000	1150.00
CARRIER_DELAY	CARRIER_DELAY	176	22.6306818	48.8769297	0	428.0000000
WEATHER_DELAY	WEATHER_DELAY	176	14.5852273	110.6717339	0	1150.00
NAS_DELAY	NAS_DELAY	176	11.4147727	16.7963808	0	102.0000000
SECURITY_DELAY	SECURITY_DELAY	176	0.1306818	1.7336902	0	23.0000000
LATE_AIRCRAFT_DELAY	LATE_AIRCRAFT_DELAY	176	14.3238636	33.5942631	0	252.0000000

The arrival delay ranges from 1 to 1,150 minutes. With such a large range in delay times could potentially mean there are both domestic airlines and international airlines within this study since domestic airlines tend to have shorter delay times compared to international airlines. This idea is further supported by both carrier delay, which ranges from 0 to 428 minutes, and weather delay, which ranges from 0 to 1,150 minutes. Carriers can be both domestic and international airlines which as mentioned before, domestic airlines tend to have shooter delay times compared to international airlines. Weather delays tend to have worse delays when traveling internationally. National Air System delay ranges from 0 to 102 minutes which means that these airlines are delayed due to airspace being too crowded at certain periods. Security delay ranges from 0 to 23 minutes which is the delay when going through security checks at airports. Late aircraft delay ranges from 0 to 252 minutes, which is the delay of airlines arriving at the airport.

Results

Looking at our correlation matrix table in Table 1, we see that only three of the five variables, carrier delay, weather delay & late aircraft delay, are significantly correlated with our dependent variable, arrival delay. From the data, we can see that carrier delay, weather delay, & late aircraft delay all have a p-value less than 0.1500. From those three variables, the weather delay is correlated strongest with arrival delay with a linear correlation coefficient of 0.892. When we take a look at the scatterplots (Table 2, 3 and 6), we see there is not a strong association from the charts, but the p-value says otherwise. The other variables such as NAS delay and Security delay have little significance to our dependent variable which we can ignore for our prediction model.

The stepwise regression method resulted in a model which includes only three of the five independent variables, because they have a p-value of less than 0.1500 which fits our minimum criteria. The final model is the following:

Arrival Delay (Minutes) = 15.28209 + 0.91601 (Carrier Delay) + 0.99451 (Weather Delay)

- **0.87742 (Late Aircraft Delay)**

The final model has an R-Square = 0.9821 (Table 8) which means that 98.21% of the variation in arrival delays are explained by carrier delay, weather delay, and late aircraft delay. A 0.9821 is a high value for R-Square which means that there is a small amount of variation which can be accounted for by some factors. The F-Statistic = 3147.82 and P-Value = <0.0001 (Table 8) is very significant.

The residual plot in Table 9 shows that there is a high amount of residuals grouped on the bottom left of the graph. The normal probability plot in Table 10 shows that points are generally close to the line, there is slight variation off the line on the low and high portion of the graph.

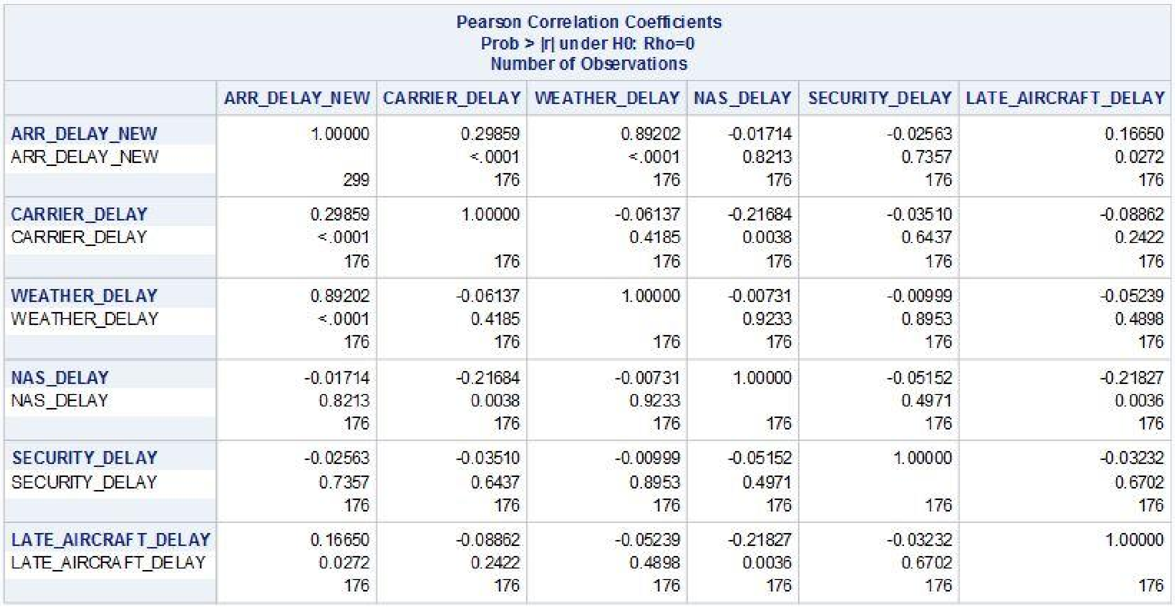
In conclusion, I am able to create a statistically significant model to predict arrival delays using carrier delay, weather delay, and late aircraft delay.

Future Work

Given that I am not aware of which airlines are used in the study whether they are domestic or international airlines or even both makes it difficult to determine what causes the arrival delays and whether this study needs to be expanded to cover specific airlines. This study only covers one year of data and may require more years of data in order for this study to be more accurate.

Appendix

Table 1 - Correlation Matrix



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	ARR_DELAY_NEW	CARRIER_DELAY	WEATHER_DELAY	NAS_DELAY	SECURITY_DELAY	LATE_AIRCRAFT_DELAY
ARR_DELAY_NEW	1.00000	0.29859	0.89202	-0.01714	-0.02563	0.16650
ARR_DELAY_NEW		<.0001	<.0001	0.8213	0.7357	0.0272
	299	176	176	176	176	176
CARRIER_DELAY	0.29859	1.00000	-0.06137	-0.21684	-0.03510	-0.08862
CARRIER_DELAY	<.0001		0.4185	0.0038	0.6437	0.2422
	176	176	176	176	176	176
WEATHER_DELAY	0.89202	-0.06137	1.00000	-0.00731	-0.00999	-0.05239
WEATHER_DELAY	<.0001	0.4185		0.9233	0.8953	0.4898
	176	176	176	176	176	176
NAS_DELAY	-0.01714	-0.21684	-0.00731	1.00000	-0.05152	-0.21827
NAS_DELAY	0.8213	0.0038	0.9233		0.4971	0.0036
	176	176	176	176	176	176
SECURITY_DELAY	-0.02563	-0.03510	-0.00999	-0.05152	1.00000	-0.03232
SECURITY_DELAY	0.7357	0.6437	0.8953	0.4971		0.6702
	176	176	176	176	176	176
LATE_AIRCRAFT_DELAY	0.16650	-0.08862	-0.05239	-0.21827	-0.03232	1.00000
LATE_AIRCRAFT_DELAY	0.0272	0.2422	0.4898	0.0036	0.6702	
	176	176	176	176	176	176

Table 2 - Scatter Plot of Arrival Delay vs Carrier Delay

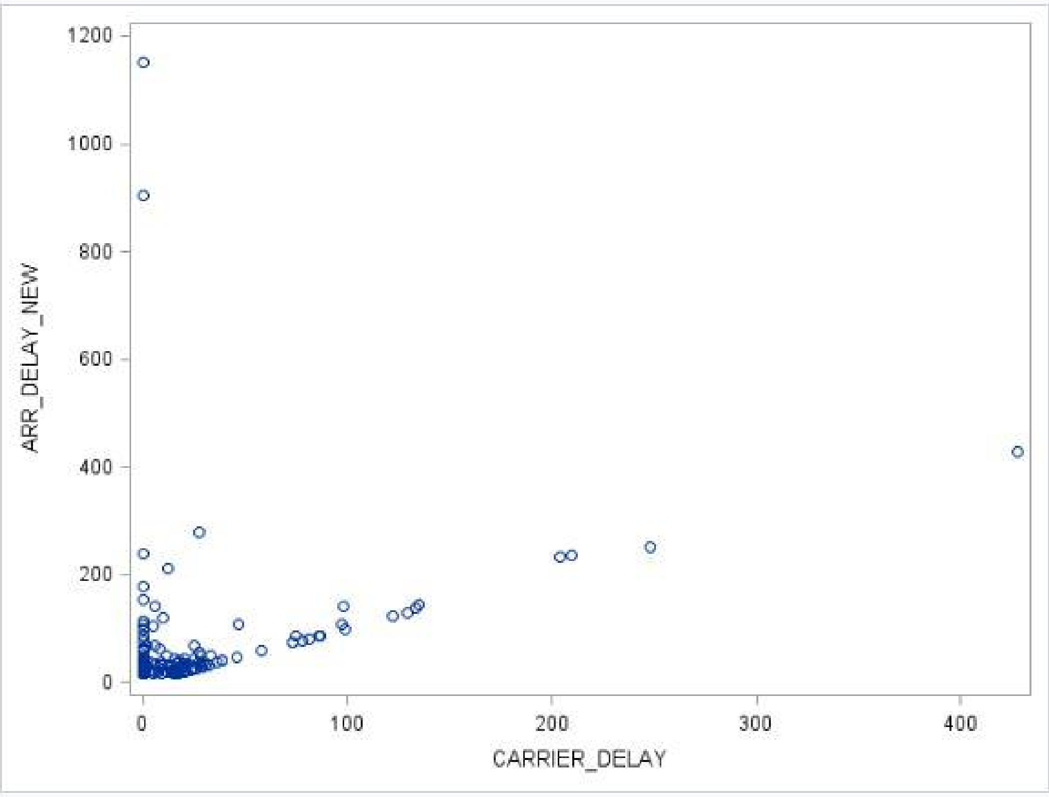


Table 3 -Scatter Plot of Arrival Delay vs Weather Delay

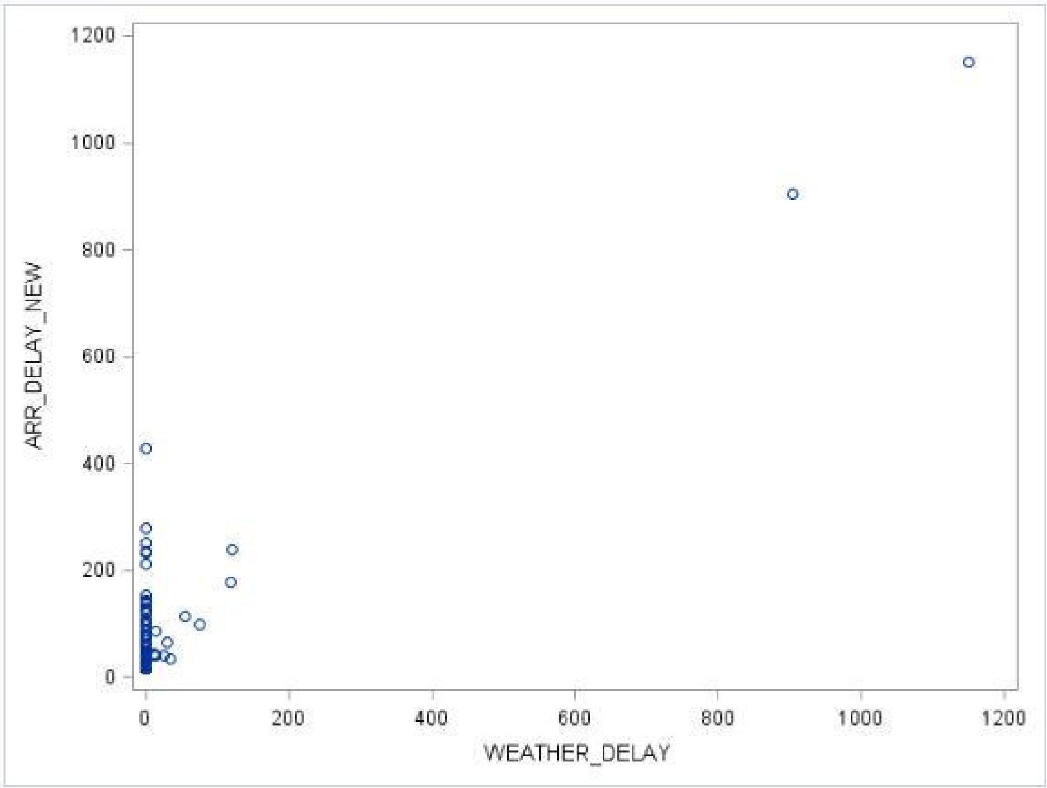


Table 4 - Scatter Plot of Arrival Delay vs NAS Delay

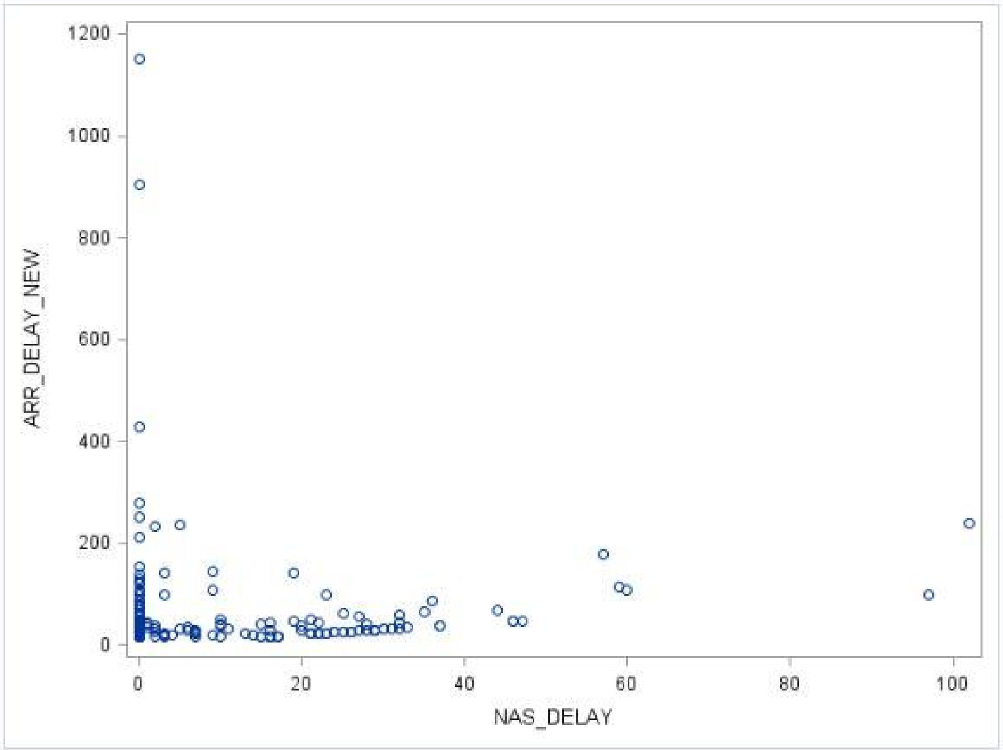


Table 5 - Scatter Plot of Arrival Delay vs Security Delay

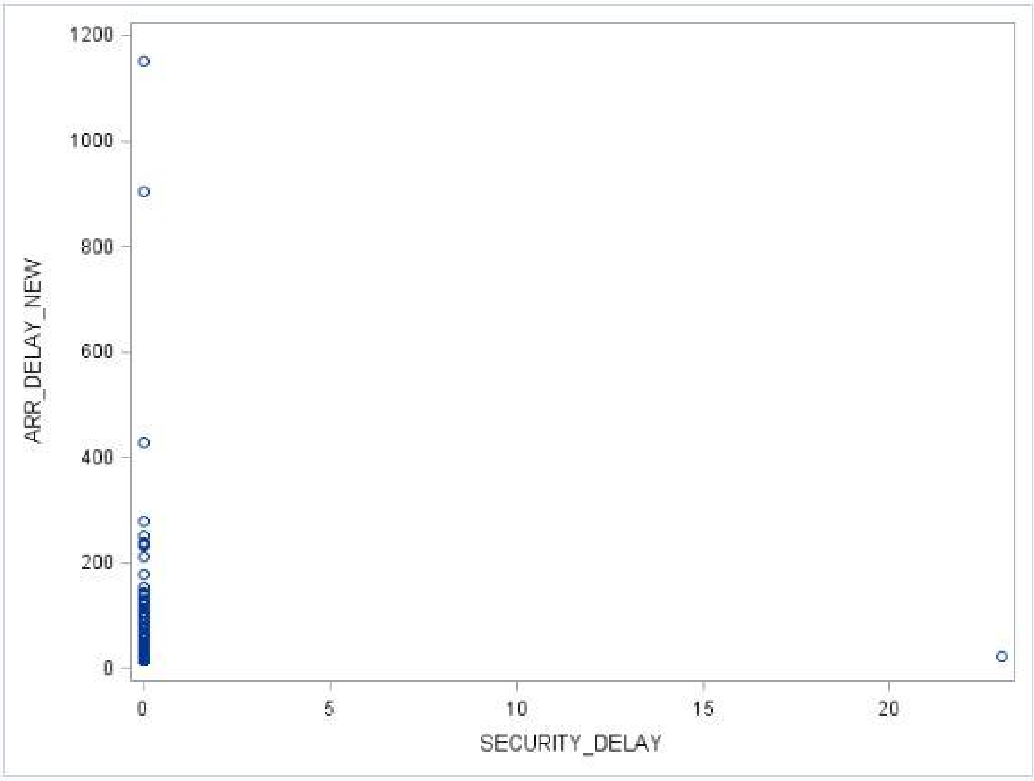


Table 6 - Scatter Plot of Arrival Delay vs Late Aircraft Delay

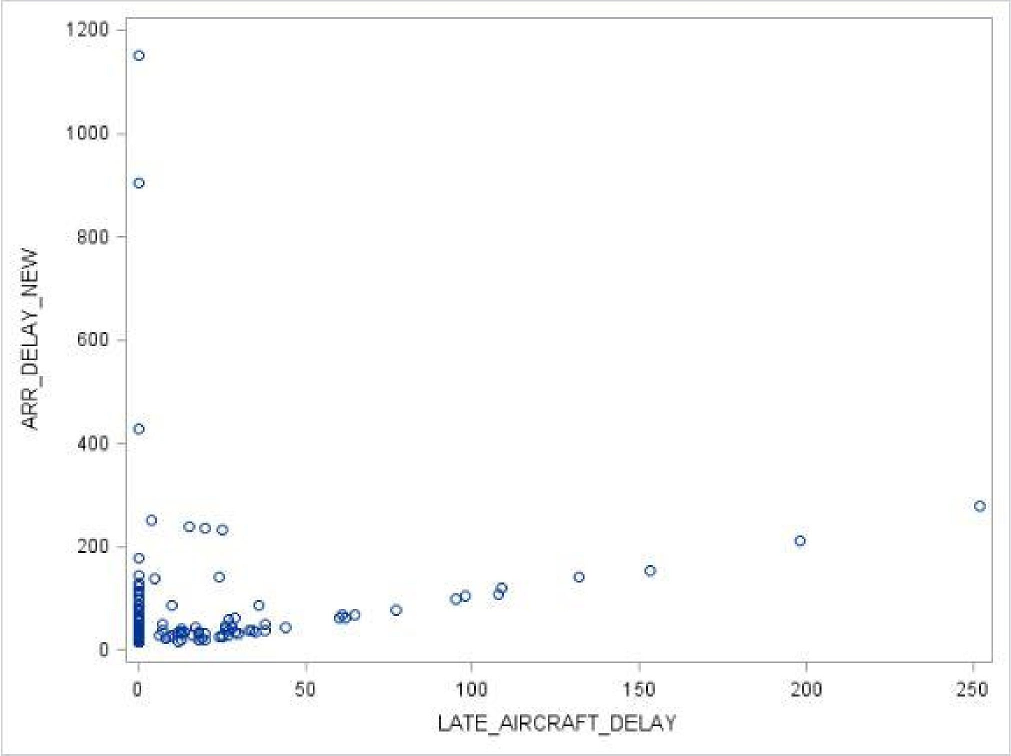
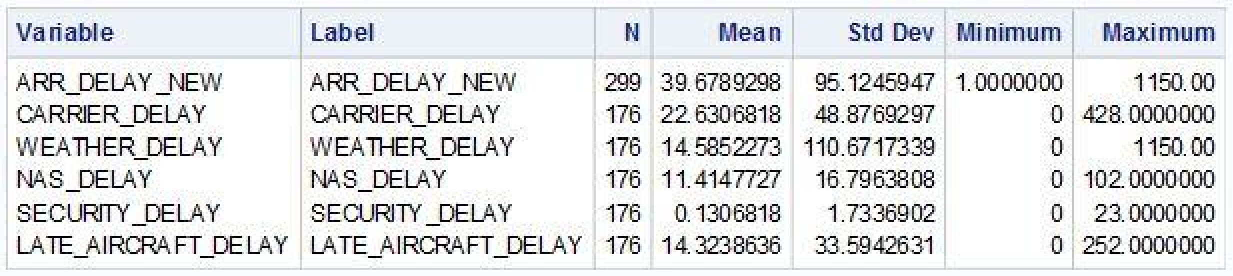
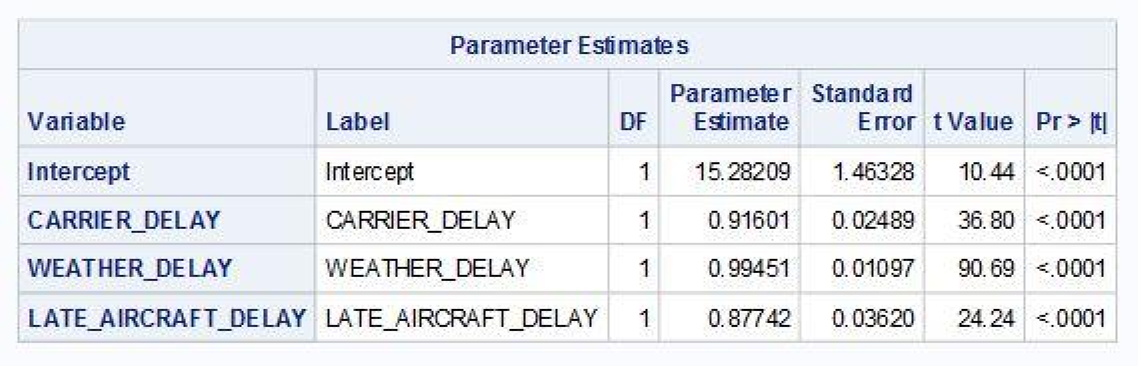


Table 7 - Descriptive Statistics



Variable	Label	N	Mean	Std Dev	Minimum	Maximum
ARR_DELAY_NEW	ARR_DELAY_NEW	299	39.6789298	95.1245947	1.0000000	1150.00
CARRIER_DELAY	CARRIER_DELAY	176	22.6306818	48.8769297	0	428.0000000
WEATHER_DELAY	WEATHER_DELAY	176	14.5852273	110.6717339	0	1150.00
NAS_DELAY	NAS_DELAY	176	11.4147727	16.7963808	0	102.0000000
SECURITY_DELAY	SECURITY_DELAY	176	0.1306818	1.7336902	0	23.0000000
LATE_AIRCRAFT_DELAY	LATE_AIRCRAFT_DELAY	176	14.3238636	33.5942631	0	252.0000000

Table 8 - Parameter Estimates



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	15.28209	1.46328	10.44	<.0001
CARRIER_DELAY	CARRIER_DELAY	1	0.91601	0.02489	36.80	<.0001
WEATHER_DELAY	WEATHER_DELAY	1	0.99451	0.01097	90.69	<.0001
LATE_AIRCRAFT_DELAY	LATE_AIRCRAFT_DELAY	1	0.87742	0.03620	24.24	<.0001

Table 9 - Residual vs Predicted Value (Arrival delay)

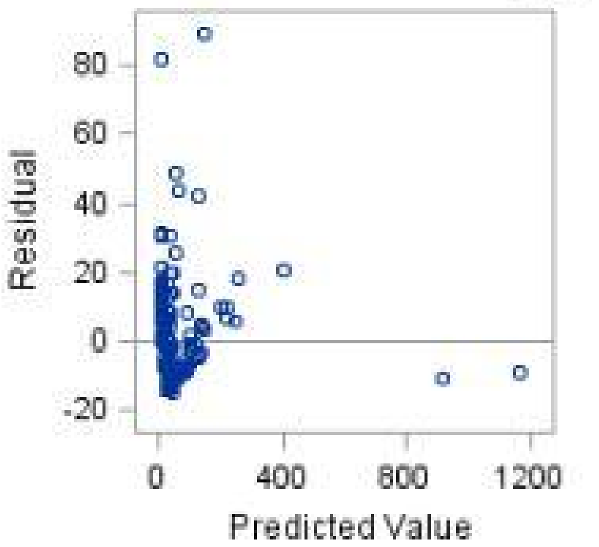
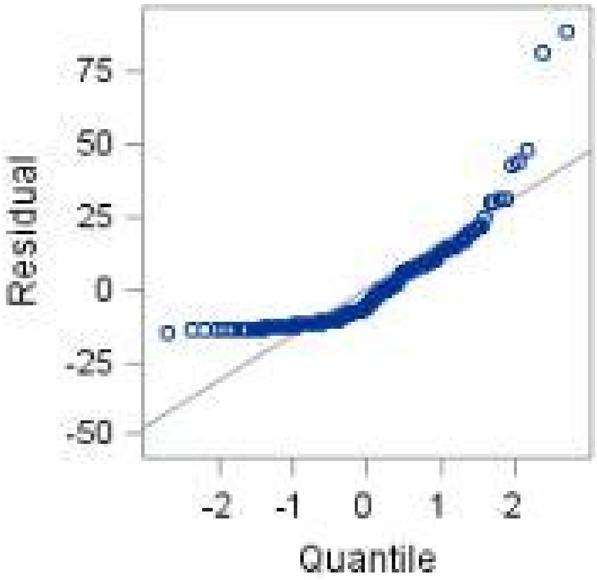


Table 10 - Normal Probability Plot


Residual
Quantile