Linear Regression Modeling of Flights to Birmingham Alabama

Abstract -

The purpose of this paper is to see which delay variable is the best at determining how long a flight will be delayed when arriving at Birmingham. When beginning this analysis, six delay variables were chosen as potential independent variables. They included departure delay, carrier delay, weather delay, national air system delay, security delay, and late aircraft delay. In order to see which variable would be the best to use a frequency distribution of all of the potential independent variables was made. When looking at the frequency distribution only departure delay didn't have a high number of values at 0 which in this case any value at 0 is a flight that had an arrival delay but the cause wasn't that variable. After this, the correlation between our dependent variable and all of the independent variables was created. When analyzing these numbers, again departure delay had the closest value to 1 and was well above the rest of the independent variables. Departure delay was then used to make a linear regression equation as it seemed to be the best independent variable out of all the options. While departure delay was the best option in this situation all of these independent variables could potentially help determine the value of the arrival delay. This makes the use of multiple regression potentially better in this case.

Exploratory Data Analysis -

Is there any way to limit the number of arrival delays at the Birmingham airport in April? This is the question that I will try to answer using linear regression. Before any analysis was started I determined which airport was the biggest in Alabama in terms of the quantity of flights. When using the table that I made, Birmingham had the most

flights out of any airport in Alabama (FIgure 1) . After determining that Birmingham was the biggest airport in Alabama I then made a smaller data set consisting of only flights that arrived in Birmingham and had an arrival delay greater than 0. This sub data set was created because the goal of this analysis is to see what the best predictor of arrival delay will be in the biggest airport in Alabama. Making the smaller data set makes it easier to answer this question.

When beginning my analysis I first determined how big the new data set was and the new data set had 537 observations and this would be all flights that had an arrival delay going into Birmingham. After this, I made histograms to see the distribution of all of the independent variables as well as our dependent variable. When graphing all of the variables, every single one was right-skewed which leads me to think that many of the delays are short and some of our independent variables may have values equal to 0(figures 2-8). This would potentially make them not that great at predicting arrival delay for example if my flight was only delayed by weather and made my flight arrive late then the value of a security delay in this situation would be 0 and wouldn't be a predictor of that arrival delay observation. To check my assumptions I made a frequency table for each variable to determine the number of observations at a certain value. When looking at the frequency table for the first potential independent variable of departure delay there aren't many values at 0 which is good because it means many of these observations could potentially help predict arrival delay. The distribution seems right skewed and there are some potential outliers as well with what seem to be delays of over 700 minutes of time delayed (figure 9). When looking at the next variables table of carrier delay it looks like left skewed with a possible outlier but, about 42 percent of the

observations are 0 which means when there was an arrival delay the variable carrier delay can't be a cause which only leaves about 58 percent of these observations to play any role in predicting arrival delay (figure 10). This is the same story when looking at the rest of the potential independent variables. They all seem to be right skewed with maybe one or two potential outliers but they all have a heavy number of observations at 0. When looking at weather delay about 92 percent of the observations are 0 with a potential outlier which also makes this variable not a predictor of arrival delay (figure 11). The next variable nas or national air system delay was also right skewed and had a high number of observations at 0 with about 56 percent and didn't seem to have any potential outliers (figure 12). Security delay had only one observation that wasn't 0 which means that security delay likely won't be a cause for arrival delay (figure 13). Lastly, late aircraft delays also had a high number of observations at 0 with about 47 percent and also didn't seem to have any potential outliers (figure 14). When looking at the distributions as well as the frequency tables I think that it's pretty clear that departure delay is the strongest predictor out of all the potential independent variables since the data is a little more spread and there are many values that are 0 or less like many of the other ones. Even after looking at all of this, I wanted to see if the low amount of weather delays was because of cancellations but only 1.086 percent of flights to Birmingham were canceled so that doesn't seem to be the case(figure 15). Another thing I wanted to look at was to see if a certain airline could be causing more delays but, when looking at the table although Southwest had the most delays they also had the most flights into Birmingham so I don't think this is the case either and all the other carriers seemed to have a similar percentage of delayed flights (figure 16). I also

looked at the departing airports to see if a potential airport could be a cause of arrival delay but, when looking at the top five airports of ATL, DFW, CLT, DEN, and ORD they all had similar percentages except Denver but, the percentage isn't too much higher than the other four so I don't this is the case either (figure17). With all of the data being right skewed I wanted to see if most of the flights had short delays so I took the percentage of flights that had a delay longer than 15 minutes and I was kind of surprised to see that 53.8 percent of flights had delays greater than 15 minutes (figure18). I also looked to see what percentage of flights came early to see how likely it is to arrive early but, only 1.086 percent of flights arrived early (figure 19).

Correlation Analysis -

For the correlation analysis, I took the same independent variables as before departure delay, carrier delay, weather delay, national air system delay, security delay, and late aircraft delay, and compared them to arrival delay. When looking at the p-values for the t-test three variables don't have significant p-values to reject the null hypothesis so they shouldn't be considered for our regression model. This leaves departure delay, carrier delay, weather delay, national air system delay, and late aircraft delay. I also wanted to see if distance was correlated with either arrival delay or departure delay but, neither p-value was significant. When looking at the coefficient correlation or r value of the remaining variables only weather delay and departure dealy are the only ones with an r value greater than .5. Weather delay has a moderately strong correlation with an r value of 0.65287 but, there weren't many observations for this variable and the r-value of departure delay is very strong further proving that this is the independent variable we should use for our model (figure 20). This is further proved

when looking at the scatterplots of all the variables because departure delay looks the best (figure 21-26).

Simple Linear Regression -

Now that the departure delay is determined to be the best independent variable out of the others I will use run a proc rec in sas to make a model. The first thing I'm looking at and interpreting is the equation. The equation in the case is arrival delay = 0.96049 * departure delay + 5.19445. This means that for every one-minute increase in departure delay the arrival delay will increase by 0.96049 minutes and when there is no departure delay the arrival delay will be 5.19445 minutes. I will next look at mean square error or MSE which is the variance in this case the number is 259.30264 and this is a really high number because ideally, you want this number to be closer to 0. Next, I will look at root mean squared error and this is the standard deviation and this number is 16.102 which is also too high and not very good. Next, I will look at the coefficient of variation which is also big with a value of 44.56424. This value is well above 10 and not very good. The last thing I'm looking at is the r^2 value which is 0.9330. This means that 93.30 percent of the variation in arrival delay can be explained by departure delay which is good (figure 27). Overall the model isn't really that good and there are a lot more factors that come into play when predicting arrival delay but, this should give an okay estimation of what the arrival may be when the departure delay is known.

Conclusion -

Overall after doing all the analysis to figure out which of the independent variables might be the best at predicting arrival delay the variable that I determined to

be the best wasn't really all that great. I think that if I were to do this analysis again I would maybe choose some more independent variables to see if I may have missed something which is entirely likely considering how the model came out. Ideally though if I was going to do this again I would want to do multiple regression because there seem to be many predictors of arrival delay. In this case, if I were to do multiple regression with the same variables I would probably choose departure delay, carrier delay, and weather delay. In conclusion, the model may not have been that great but, it will still give an okay idea of what the arrival time might be.

Appendix

Figure 1

	DEST	flights
1	BHM	1288
2	HSV	675
3	MOB	205
4	MGM	197
5	DHN	60

Figure 2

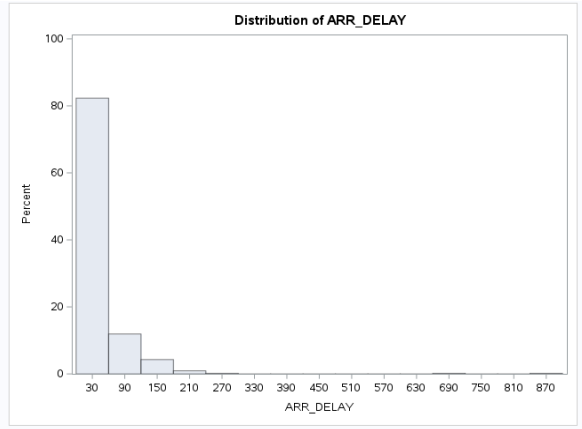

Figure 3

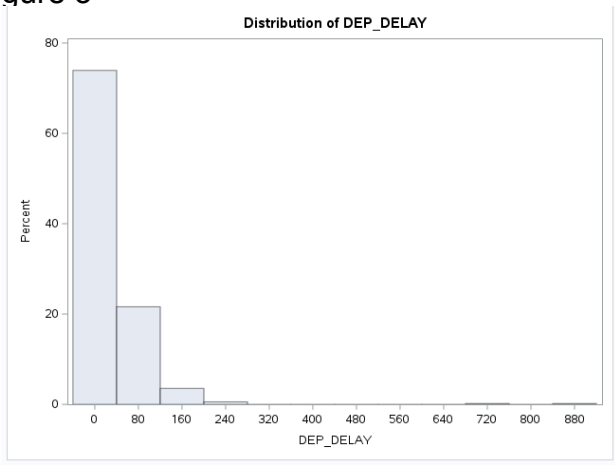

Figure 4

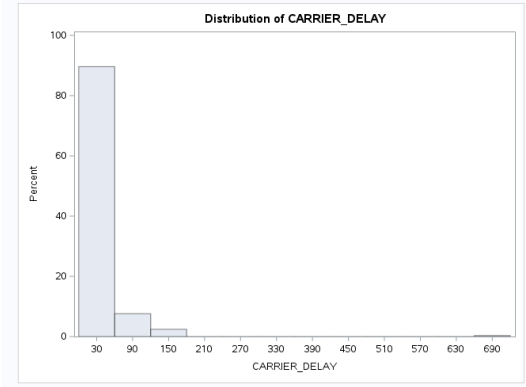

Figure 5

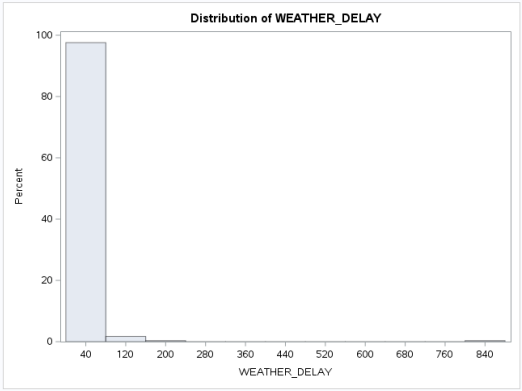

Figure 6

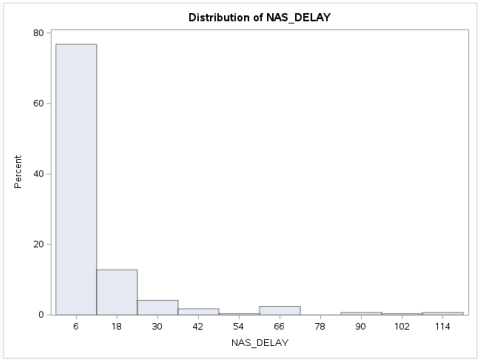

Figure 7

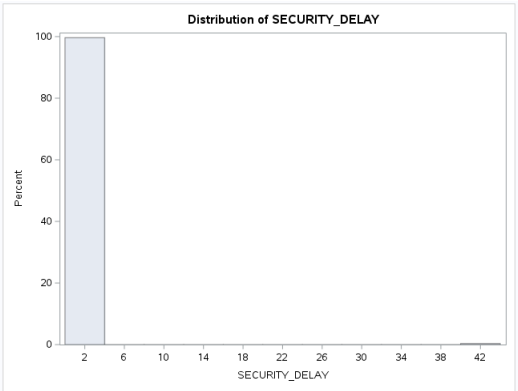

Figure 8

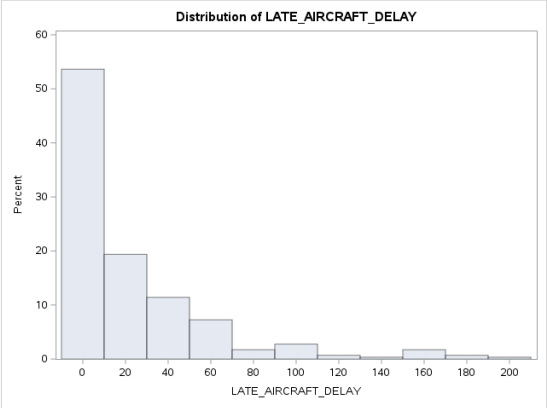

Figure 9

DEP_DELAY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-12	1	0.19	1	0.19
-11	1	0.19	2	0.37
-10	1	0.19	3	0.56
-9	4	0.74	7	1.30
-8	6	1.12	13	2.42
-7	6	1.12	19	3.54
-6	9	1.68	28	5.21
-5	10	1.86	38	7.08
-4	12	2.23	50	9.31
-3	26	4.84	76	14.15
-2	19	3.54	95	17.69
-1	22	4.10	117	21.79
0	12	2.23	129	24.02
1	5	0.93	134	24.95
2	13	2.42	147	27.37
3	5	0.93	152	28.31
4	10	1.86	162	30.17
5	9	1.68	171	31.84
6	9	1.68	180	33.52
7	14	2.61	194	36.13

Figure 10

CARRIER_DELAY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	121	41.87	121	41.87
1	3	1.04	124	42.91
2	9	3.11	133	46.02
3	4	1.38	137	47.40
4	4	1.38	141	48.79
5	5	1.73	146	50.52
6	1	0.35	147	50.87
7	3	1.04	150	51.90
8	4	1.38	154	53.29
9	8	2.77	162	56.06
11	4	1.38	166	57.44
13	1	0.35	167	57.79
14	2	0.69	169	58.48
15	2	0.69	171	59.17
16	3	1.04	174	60.21
17	4	1.38	178	61.59
18	7	2.42	185	64.01
19	4	1.38	189	65.40
20	5	1.73	194	67.13

Figure 11

WEATHER_DELAY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	267	92.39	267	92.39
4	1	0.35	268	92.73
5	1	0.35	269	93.08
8	1	0.35	270	93.43
9	1	0.35	271	93.77
10	1	0.35	272	94.12
12	1	0.35	273	94.46
26	1	0.35	274	94.81
31	1	0.35	275	95.16
34	1	0.35	276	95.50
37	1	0.35	277	95.85
39	1	0.35	278	96.19
46	1	0.35	279	96.54
53	1	0.35	280	96.89
57	1	0.35	281	97.23
69	1	0.35	282	97.58
80	1	0.35	283	97.92
93	1	0.35	284	98.27
116	1	0.35	285	98.62
122	1	0.35	286	98.96
135	1	0.35	287	99.31
179	1	0.35	288	99.65
876	1	0.35	289	100.00

Figure 12

NAS_DELAY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	162	56.06	162	56.06
1	10	3.46	172	59.52
2	6	2.08	178	61.59
3	2	0.69	180	62.28
4	9	3.11	189	65.40
5	7	2.42	196	67.82
6	5	1.73	201	69.55
7	2	0.69	203	70.24
8	8	2.77	211	73.01
9	5	1.73	216	74.74
10	4	1.38	220	76.12
11	2	0.69	222	76.82
12	4	1.38	226	78.20
13	3	1.04	229	79.24
14	1	0.35	230	79.58
15	7	2.42	237	82.01
16	3	1.04	240	83.04
17	3	1.04	243	84.08
18	4	1.38	247	85.47
19	4	1.38	251	86.85
20	1	0.35	252	87.20

Figure 13

SECURITY_DELAY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	288	99.65	288	99.65
42	1	0.35	289	100.00
Frequency Missing = 248				

Figure 14

LATE_AIRCRAFT_DELAY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	137	47.40	137	47.40
1	1	0.35	138	47.75
2	1	0.35	139	48.10
3	2	0.69	141	48.79
4	6	2.08	147	50.87
6	3	1.04	150	51.90
7	2	0.69	152	52.60
8	2	0.69	154	53.29
9	1	0.35	155	53.63
11	1	0.35	156	53.98
12	2	0.69	158	54.67
13	3	1.04	161	55.71
14	5	1.73	166	57.44
15	3	1.04	169	58.48
16	5	1.73	174	60.21
17	2	0.69	176	60.90
18	5	1.73	181	62.63
19	1	0.35	182	62.98
20	4	1.38	186	64.36

Figure 15

percentofcancelledflights
1.086957

Figure 16

	OP_UNIQUE_CARRIER	airlinecount	delayedflights	percentagedelayed
1	WN	317	173	54.574132492
2	OH	309	95	30.74433657
3	DL	222	80	36.036036036
4	OO	181	80	44.198895028
5	AA	108	46	42.592592593
6	MQ	79	41	51.898734177
7	9E	72	22	30.555555556

Figure 17

	ORIGIN	airportcount	delayedcount	percentagedelayed
1	ATL	222	80	36.036036036
2	DFW	170	62	36.470588235
3	CLT	145	53	36.551724138
4	DEN	90	51	56.666666667
5	ORD	86	33	38.372093023

Figure 18

percentofgreater15delay
53.8175

Figure 19

percentofearlyflights
1.086957

Figure 20

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	ARR_DELAY	DEP_DELAY
CARRIER_DELAY	0.49524 <.0001 289	0.51295 <.0001 289
WEATHER_DELAY	0.65287 <.0001 289	0.64606 <.0001 289
NAS_DELAY	0.16274 0.0056 289	-0.04470 0.4491 289
SECURITY_DELAY	-0.00508 0.9315 289	0.02695 0.6482 289
LATE_AIRCRAFT_DELAY	0.32401 <.0001 289	0.33731 <.0001 289
DEP_DELAY	0.96593 <.0001 537	1.00000 537
DISTANCE	0.02129 0.6225 537	0.00738 0.8644 537

Figure 21

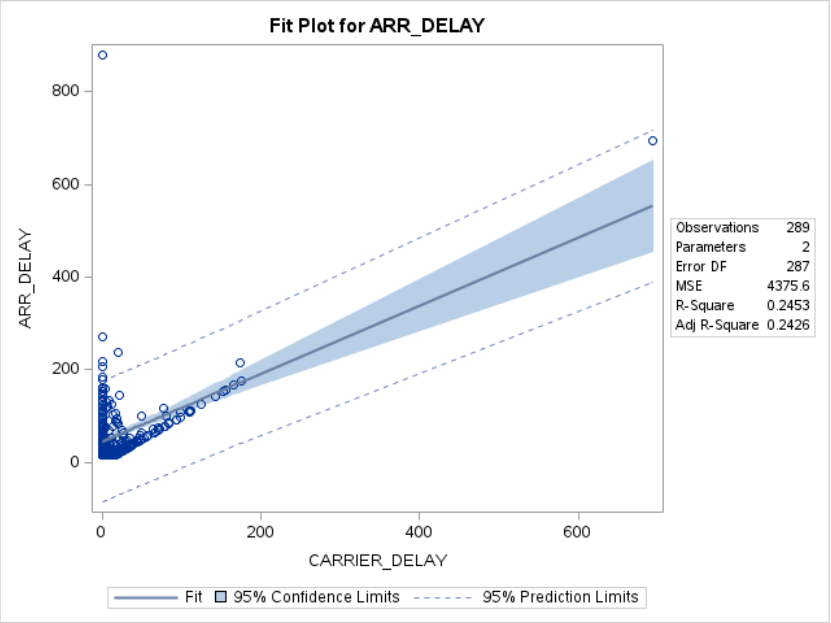

Figure 22

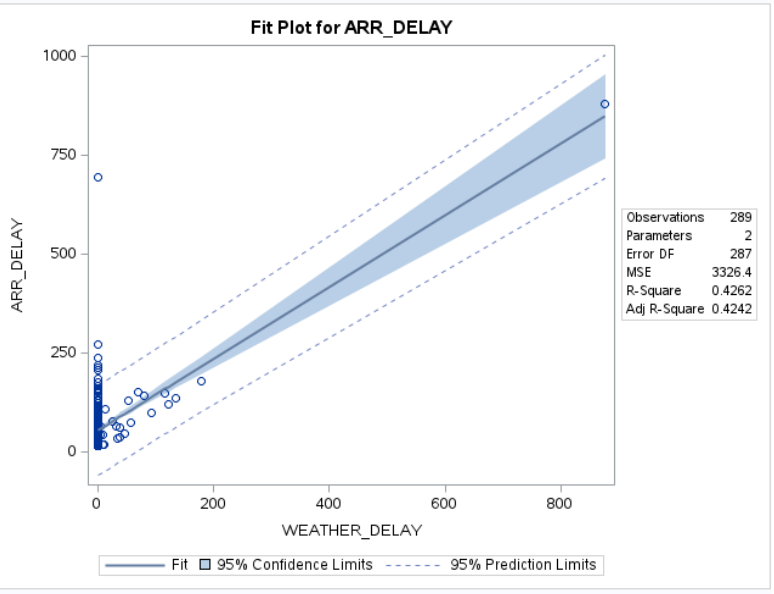

Figure 23

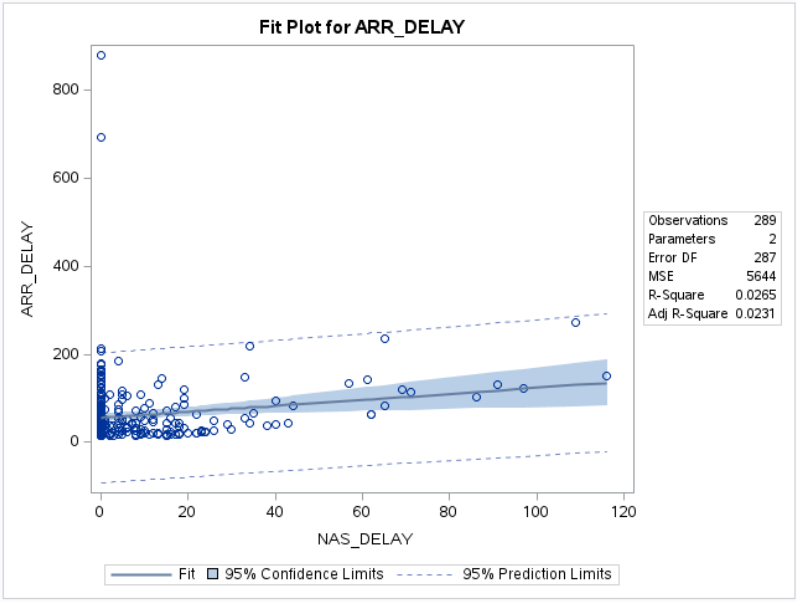

Figure 24

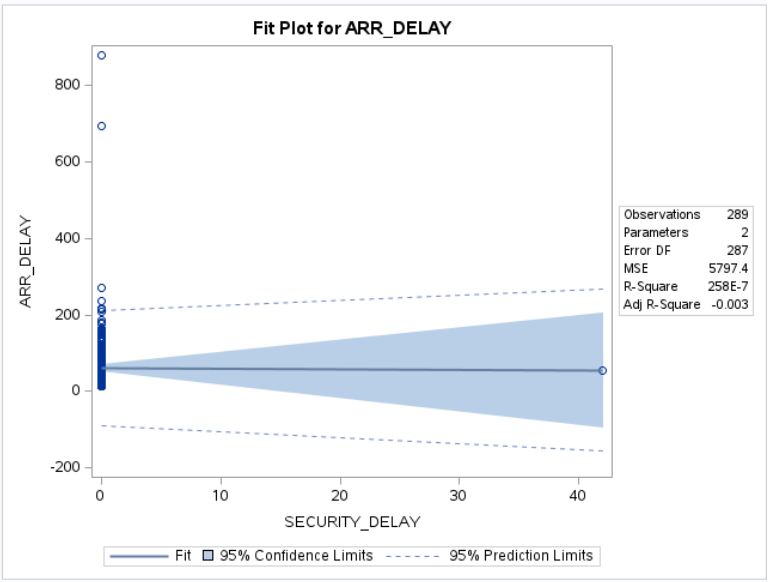

Figure 25

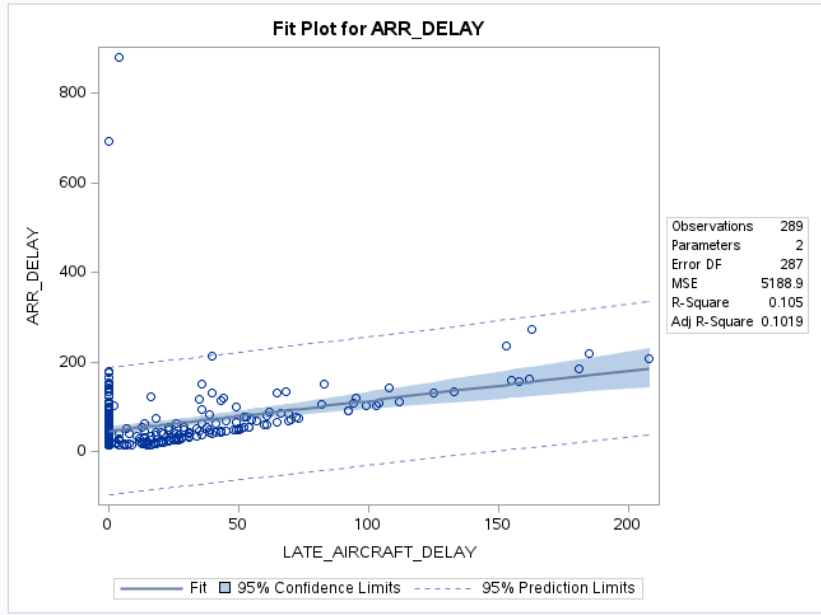

Figure 26

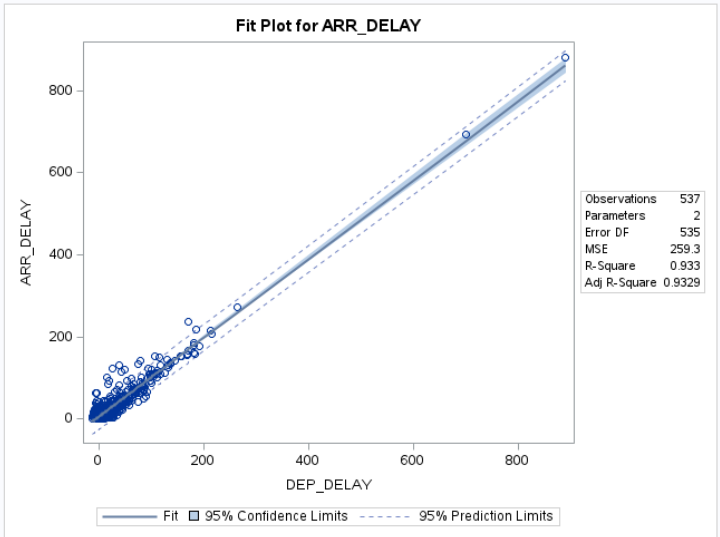

Figure 27

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1932539	1932539	7452.83	<.0001
Error	535	138727	259.30264		
Corrected Total	536	2071266			

Root MSE	16.10288	R-Square	0.9330
Dependent Mean	36.13408	Adj R-Sq	0.9329
Coeff Var	44.56424		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.19445	0.78187	6.64	<.0001
DEP_DELAY	1	0.96049	0.01113	86.33	<.0001

Code:

```
/* Generated Code (IMPORT) */
/* Source File: T_ONTIME_REPORTING.csv */
/* Source Path: /home/u63542550/sasuser.v94 */
/* Code generated on: 9/13/23, 5:58 PM */

%web_drop_table(WORK.IMPORT);


FILENAME REFFILE '/home/u63542550/sasuser.v94/T_ONTIME_REPORTING.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.AlabamaFlights;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.AlabamaFlights; RUN;


%web_open_table(WORK.IMPORT);

proc print data=alabamaflights(obs=10);
run;

/*find biggest airport */
proc sql;
	create table citycount as
	select DEST, count(*) as flights
	from AlabamaFlights
	where DEST_STATE_ABR = 'AL'
	group by DEST
	order by flights desc;
quit;

/*making new dataset where the dest is birmingham and arrival delay >0*/
data birminghamflights;
	set alabamaflights;
	if DEST = 'BHM' and arr_delay >0;
run;

/*dataset with dest as birmingham*/
data allbirminghamflights;
	set alabamaflights;
	if DEST = 'BHM';
run;
```

```
PROC CONTENTS DATA=WORK.birminghamFlights; RUN;

/* graphing all variables*/
proc univariate data=birminghamflights;
  var arr_delay dep_delay carrier_delay weather_delay nas_delay security_delay
late_aircraft_delay;
  histogram arr_delay dep_delay carrier_delay weather_delay nas_delay security_delay
late_aircraft_delay;
  run;
/* frequency table of all variables*/
proc freq data=birminghamflights;
  tables arr_delay dep_delay carrier_delay weather_delay nas_delay security_delay
late_aircraft_delay;
run;
/* percentage of cancelled flights*/
proc sql;
select (CANCELLED/TOTALFLIGHTS)*100 as percentofcancelledflights
from (select count(*) as CANCELLED from allbirminghamflights where
CANCELLED=1),
(select count(*) as TOTALFLIGHTS from allbirminghamflights) as counts;
quit;

/* airline by flights */
proc sql;
   create table flightsByAirlines as
   select OP_UNIQUE_CARRIER,
         count(*) as airlinecount,
         sum(arr_delay > 0) as delayedflights,
         (sum(arr_delay > 0)/count(*))*100 as percentagedelayed
   from allbirminghamflights
   where DEST = 'BHM'
   group by OP_UNIQUE_CARRIER
   order by airlinecount desc;
quit;

/* Produce a histogram for the “Departure Delay” and note if there were flights that
would be considered
unusual values. */
proc univariate data=allbirminghamflights;
  histogram Dep_delay;
run;

/* Create a frequency table that identifies the top 5 airports where flights originated from
that arrived in
your city in April. What airline had the most flights to your city in April? */
```

```
proc sql;
   create table topFiveFrequency as
   select ORIGIN,
         count(*) as airportcount,
         sum(arr_delay > 0) as delayedcount,
        (sum(arr_delay > 0)/count(*))*100 as percentagedelayed
   from allbirminghamflights
   where DEST = 'BHM'
   group by ORIGIN
   order by airportcount desc;
quit;

/* What percentage of flights arrived early (arrival time < 0)? */
proc sql;
select (ARR_TIME/TOTALFLIGHTS)*100 as percentofearlyflights
from (select count(*) as ARR_TIME from allbirminghamflights where ARR_TIME<0),
(select count(*) as TOTALFLIGHTS from allbirminghamflights) as counts;
quit;

/*Of all the delayed flights, what percentage of flights had a 15 minute or longer delay?*/
proc sql;
select (arr_DELAY/TOTALFLIGHTS)*100 as percentofgreater15delay
from (select count(*) as arr_DELAY from birminghamflights where arr_DELAY>=15),
(select count(*) as TOTALFLIGHTS from birminghamflights) as counts;
quit;

/* proc corr to show correlation between are independant variable choices and
dependant variable */
proc corr data=birminghamflights;
        var arr_delay dep_delay;
        with carrier_delay weather_delay nas_delay security_delay late_aircraft_delay
dep_delay distance;
        run;

/* proc reg for all delay variables*/
proc reg data=birminghamflights;
        model ARR_DELAY=Carrier_delay;
        run;

proc reg data=birminghamflights;
        model ARR_DELAY=Weather_delay;
        run;

proc reg data=birminghamflights;
        model ARR_DELAY=Nas_delay;
```

```
        run;

proc reg data=birminghamflights;
        model ARR_DELAY=Security_delay;
        run;

proc reg data=birminghamflights;
        model ARR_DELAY=late_aircraft_delay;
        run;

proc reg data=birminghamflights;
        model ARR_DELAY=dep_delay;
        run;
```