# Airlines Delay at major Airport in the United States

Submitted by:

Pranab Bhadani

Gaurav Derasaria

Diksha Yadav

***Introduction to the Problem***

United States of America is home to seventeen of the world’s top thirty busiest airports in the world. There has been an incessant growth in the number of flights over the years. Today, approximately 11 million flight departures take place every year [1]. The sheer size of the country makes air travel one of the most viable modes of transport for longer distance. This has applied a significant pressure on the airline industry to deliver in terms of safety and time. While safety in the airlines industry has reached a satisfactory level, the same cannot be said about the time delays.

Given the number of passengers travelling daily, delays in airline can impact a significant amount of people. This leaves very low margin for error for both airlines and the airport authorities. Unfortunately, according to the Bureau of transportation statistics 22.26% of the flights arrived more than 15 mins late for the year 2012. We attempt to study what are the major factors that lead to the regular delay in flights.

Based on the past studies on flight delays, we believe that location of the airports, Distance of travel, arrival time, departure time and air traffic are major factors in delay of the airport. Also, we believe that factors like the airline company and the effect of harsh weather in winter can have a significant effect on the air traffic delay. Based on the followings factors we propose three models that explore the impact of factors on aircraft delay and their respective hypothesis:

1. What are the significant parameters that affect the delays in flight throughout the year? We use the data for the flights between 11 busiest airports in the United States for the year 2008.We hypothesize that along that with the factors that were included in the past studies winter and airline companies have a significant impact on the delay of flights. We appreciate the complications that could surface due to the introduction of dummy variables in multiple categories. Our aim here is to examine the impacts of various factors on the delay of flights.
2. Second, how is the flight delay affected over time for an airline? We use a time series data of 4 years (2013-2016) to study the change in delay of an airline. The motivation behind this model is to judge the performance of the airline based on the delay and try to infer the popularity of the airline. We hypothesize that there is the change is performance of the airlines over the years and time has a significant effect on the delay.
3. Third, we seek to study the variation of flight delays through a significant event. We study the delay of aircrafts from August 2001 through February 2002. The data is placed around the 9/11 attacks. Two months after 9/11 attack took place in U.S. airports, airport security was federalized and the Aviation and Transportation Security Act was passed. Prior to 9/11, airports handled security. We hypothesize that there would be an increase in time delay post 9/11 attacks. We appreciate the fact that some of the increase in the delay could be due to seasonal variation.

We believe that an Ordinary Least Square is the appropriate model in this case. The model will help us to check the statistical significance of various factors and thus, help us determine if the factor affects the delay in any form. We seek the impact on delay across various cross-sections and time series.

**Model**

The inspiration behind our models is to seek an in depth understanding of flight delay and study what factors have the greatest impact. There have been previous studies on the financial implications of delays.

One such significant study was the “Total Delay Impact Study” prepared by Michael Bell and others. The study focused on analyzing flight delay data from 2007 to calculate the economic impact on both airlines and passengers, including the cost of lost demand and the collective impact of these costs on the U.S. economy. The study found that airlines with high rates of delay also have higher operating costs overall. The $8.3 billion direct cost to airlines included increased expenses for crew, fuel and maintenance, among others. The new study emphasizes that not all delays can or should be eliminated, especially those due to mechanical failures and severe weather that are necessary to protect passenger safety but, increased scientific knowledge can improve operations and influence policy.

Our goal of the models is to extend the above study and instead of finding the relation of flight delay on the economic effect, we wanted to look for the main causes which leads to airline delay. Our aim here is to provide an insight that could in turn help reduce the financial implications by designing a better system.

We have segregated our entire study in three parts, each of which is looking on the impact of the airline delay from three different angles.

In first analysis, we collected data of the eleven busiest airports in United States for the year 2008.Our dependent variable is arrival delay. The first three independent variables are TaxiIn, Taxiout and Speed. TaxiIn and Taxiout try to mirror the effects of delay due to congestion and traffic. Few of these airports have been notorious for delays. The introduction of a total of twenty dummy variables(10 each) for origin and destination aimed to capture such impacts. We introduced four dummies for five most popular airlines in the United States. Each individual airline would have varying impacts on delay. Next, we account for harsh weather generally experienced during winters. Thus, we have a dummy for winter and non-winter seasons. We have introduced 3 dummies for the time of arrival and departure of flights by dividing the 24 hours of the day into 4 parts (00:00 to 6:00, 6:00 to 12:00 , 12:00 to 18:00, 18:00 to 00:00) . The time of arrival and departure could have a significant impact on the model. Finally, we have dummies for 6 of the days of a week. We understand that while interpreting the model intercept is of least importance due to many dummies.

Second analysis looks to see the yearly trend of the air delay for an air carrier. The motivation behind gauzing the trend of the flight delay, spanning across various years is to seek the performance of an airline over time.

Since we have a data of flight delays spanning across 2013 to 2016, we can easily apply regression with proper use of dummy variables to study the trend of flight delays across air carriers and across airports.

Since the cross section i.e., airline in the sample data across various time is constant, thus it qualifies as a panel data. We can apply fix-effect model to study the trend mentioned above. Fixing effect across carriers and using dummy variables to represent years 2013 to 2016.

So, in this regression model, we have taken air delay as our dependent variable and have used percent of aircraft delayed due to carrier effect delay, national aviation system delay, weather delay and security delay as our independent variable. Now the yearly analysis from 2013 to 2016 is a time series analysis for which we have introduced three dummy variables whose beta estimates gives us the air delay trends across years.

The result from the studied model indicate that all the independent variables turn out to be significant and we can see from the result that there is a specific trend across the year from 2013 to 2016 based on the beta values of the dummy variables. For the Alaska Airlines, the delay is minimum in the year 2013 and which substantially increases with time.

We have also found that there is multicollinearity issue between percentage of flight delay due to airline and late\_carrier\_ct, we then decided to remove the variable late\_carrier\_ct since it doesn’t have much impact on the Rsquare (minimal reduce of 0.98 to 0.97).

Third Analysis

Two months after 9/11 attack took place in U.S. airports, airport security was federalized and the [Aviation and Transportation Security Act](http://www.gpo.gov/fdsys/pkg/PLAW-107publ71/content-detail.html) was passed. Prior to 9/11, security was handled by each airport separately.

The new act implemented procedures which included stricter guidelines on passenger and luggage screening. For example, some machinery and procedures were introduced to scan for weapons and destructive items, removing shoes and banning liquids. So, clearly the time for security check increased after 9/11.

The immediate effect of attack on airlines delay can be seen best by taking the data of airline delays before 9/11 and after the security act was passed. So, we have run regression model on data from August 2001 to February 2002. We have made dummy variables for each month beginning from September 2001 to February 2002. The delay in August 2001 is taken care of by the intercept. We took data for only New York State, i.e flights which have either their origin or destination as one of the airports of New York State. We did this because we were inquisitive to find that whether the huge impact that 9/11 attack had on New York State airlines can be verified statistically. After running out model we have tried to verify our expected hypothesis of significant delay changes at New York airports after the attack using parameter estimates, standardized estimates and p-value of dummy variables of included months in the output.

All the independent dummy variables except September came as significant because the September month showed a truly different trend of delay than other months. There was a significant delay in month of September while in other months the changes are significant but it is still less than what we observed in September.

Months of November and December also showed a little rise in delay which we think can be contributed to the strict security procedures accepted under TSA after 2 months of 9/11 attack.

After November and December we didn’t find any different trend from the usual delay.

This trend can be seen best by taking the data of airline delays before 9/11 and after the security act was passed. So, we have run regression model on data from August 2001 to February 2002.

**Data Description and Data Source:**

For this project, we have used publicly available data from Bureau of Transportation Statistics to analyze and predict flight departure delays for a set of commercial flights across multiple top airports in united states.

The original data set contains information for all the commercial flights in US from the year 1987 to 2016. Because of the extremely large size of the data set we have extract a reasonable subset of data for our analysis. The details of the data are as follows:

1. *For the regression explained in the first category*

* 11 major airport across US: based on the busyness of the airport
* 5 major airlines.
* Flight Features: These include the departure time, arrival time, distance covered by the flight, Air time, origin airport, destination airport, month.
* One year: 2008

1. *For the regression explained in the second category*

* All major airport across US present in the data set: based on the business of the airport
* 5 major airlines.
* Features: These include the carrier\_ct(percentage of delayed flight due to airlines) , nas\_ct (percentage of delayed flight due to national aviation security) , weather\_ct (percentage of delayed flight due to bad weather) , security\_ct (percentage of flight delay due to security check)
* Year: 2013 to 2016

1. *For the regression explained in the second category*

* All the airport throughout New York State present in the data set: based on the business of the airport
* 5 major airlines.
* Features: Distance between the airports

Year: Aug 2001 to Feb 2002

**Potential Specification Error:**

1. Connecting flight effect: We do not have the parameters that would measure the delay which could be introduced in case of the connecting flights.
2. Air Carrier specification**:** Airplanes specification like size, model, age of the planes can also be a factor impacting in the total delay of the planes.
3. We don’t have much information about the airport features like size of the terminal, boarding and de-boarding time in airports. These features could play an important role in impacting time delay.

**Regression Analysis and Result**

**Part-1**

Our final model is:

Where ,

Y = arrdelay, in minutes

X2 = TaxiIn **time** The difference between the Wheels On time and Gate In time, in minutes.

X*3* = TaxiOut time spent by a flight between its actual off-block**time** (AOBT) and actual take-off **time** (ATOT), in minutes

X*4* = Speed in miles/min

D*1-* D*10* = Dummy variable for Origin of the airport

(‘ATL’ ‘LAX’ ‘ORD’ ‘DFW’ ‘JFK’ ‘DEN’ ‘SFO’ ‘CLT’ ‘LAS’ ‘IAH’ )

D*11-* D*20* = Dummy variable for Destination of the airport

(‘ATL’ ‘LAX’ ‘ORD’ ‘DFW’ ‘JFK’ ‘DEN’ ‘SFO’ ‘CLT’ ‘LAS’ ‘IAH’ )

D*21-* D*26* = Dummy for day of the week

D*27* = Dummy for Winter

D*28-* D*30*  = Dummies for the Arrival hours

(00:00 to 6:00, 6:00  to 12:00 , 12:00 to 18:00, 18:00 to 00:00)

D*31-* D33 = Dummies for the Departure hours

(00:00 to 6:00, 6:00  to 12:00 , 12:00 to 18:00, 18:00 to 00:00)

D*34-* D*37*  = Dummies for the airline company

The final regression equation of the analysis is stastically significant with an F value of 575.02 and a p value of less than 0.0001. Thus, the model is significant and we accept the claim that the there is a significant relationship between arrdelay and independent variable set.

All 3 continous variables have high t values and are statistically significant at the 0.05 level of significance. Amongst the three categorical variables taxiout seems to have the most significant impact that can be observed from the standardized estimate of 0.20059. The taxiout time is a variable that mirrors the congestion and traffic at the airport. Thus, a higher Taxiout time would result in a higher delay. The taxxin time variable seems to have a similar trend but a comparatively less significant impact due to a comparatively lower t value and standardized estimate.

Amongst, the dummy variable introduced to account for various changes across cross-sections it is observed that if the flight is arriving to or departing from Chicago airport the delay is bound to have the highest impact. This can be observed from high standardized value of 0.03532 compared to other origin and destination airports. The day3(Wednesday) can be thought to have the least delay due to significant standardized estimate of -0.03741 as compared to the standardized estimates of the other days. Sunday could lead to the most amount of delay since the standardized estimates of all other days are negative indicating that the day that was left out faces the problem of highest delays.

Winter causes a higher delay observed from the value of 0.03481 for the standardized estimate. 00:00 to 06:00 hours has the highest impact of flight delay due to a high parameter estimate of 33.1 and the highest positive standardized estimates amongst all the dummy variables. American Airlines and United Airlines seem to have a higher impact on delay than the other airlines due to a high positive standardized estimate of 0.03668 and 0.03446 respectively.

**Part-2**

Y = B0 + B1\*X1 + B2\*X2 + B3\*X3 + B4\*X4 + B5\*D1 + B6\* D2 + B7\*D3 + U-hat

X1: CARRIER\_CT (percent of aircraft delayed due to carrier effect)

X2: WEATHER\_CT (percent of aircraft delayed due to weather effect)

X3: NAS\_CT (percent of aircraft delayed due National Aviation system effect)

X4: SECURITY\_CT (percent of aircraft delayed due to security effect)

D1: dummy variable for the year 2016

D2: dummy variable for the year 2015

D3: dummy variable for the year 2014

B0: intercept that take care of the value in 2013.

Y = -245.31 + 94.49661\*X1 + 236.30\*X2 + 52.94171\* X3 + 25.754 \* X4 + 80.70629 \* D1 + 97.641 \* D2 + 102.76 \* D3 + u

The final regression model is statistically significant, as F value is 13387.7 has a corresponding value of <0.0001. Therefore, model shows a strong relation between dependent and the independent variables i.e. the relation between the arrival time delay and the carrier\_ct, weather\_ct, NAS\_ct, Security\_ct.

Table1 indicates that all independent variables are significant in defining the relationship between independent and the dependent variables at 0.05 alpha level. We can say with 95% level of confidence that a relationship exists between the dependent and all the independent variables.

The adjusted r-squared term 0.9705 measures the goodness of fit of the regression plane to the data. This suggest that around 97% of the variation in the arrival delay is explained by the all the independent variables in the mentioned equation.

The coefficient of variation, 46.33045 compares the mean standard error for the entire regression. Out of all the independent variables security delay causes the maximum impact on the overall arrival delay of the Alaska Airlines.

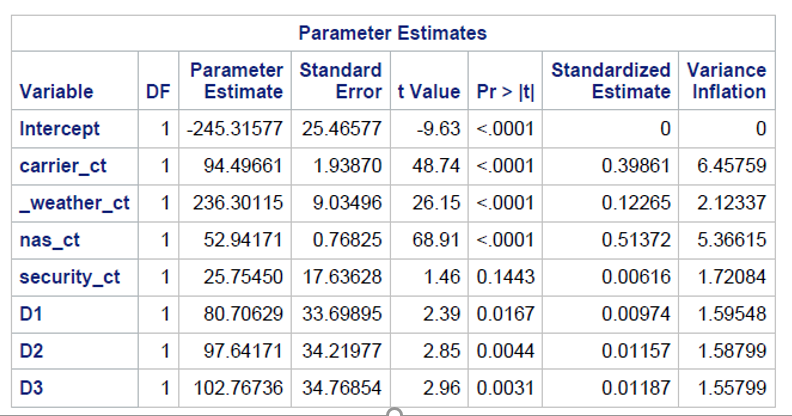


Table 1

**Part-3**

**Regression Equation**

Dep\_delay=  β1+ β2Sep01+ β3Oct01+ β4Nov01+ β5Dec01+ β6Jan02+ β7Feb02 + β8distance +u

**Definitions**

arrdelay=arrival delay

Sep01=Dummy variable for September 2001 month

Oct01=Dummy variable for October 2001 month

Nov01=Dummy variable for November 2001 month

Dec01=Dummy variable for December 2001 month

Jan02=Dummy variable for January 2002 month

Feb02=Dummy variable for February 2002 month

Distance=distance covered by the flight

Dep\_delay=  7.0997+ 0.0978Sep01-3.1636Oct01-2.7177Nov01 -2.6485Dec01-3.4963Jan02-4.8261Feb02 +0.00182distance+ u

The final regression model is statistically significant, as F value is 292.30 has a corresponding value of <0.001. Therefore, model shows a strong relation between dependent and the independent variables.

**CONCLUSION**

The regression analysis results support our initial hypothesis set that there is a trend in airline delay in relation to airport, carrier, seasonal variation, day of the week etc. Also, the impact of 9/11 on departure delay of airlines can be clearly seen from our regression results. The results of 9/11 regression analysis clearly prove that there was a sudden increase in departure delay in September 2001 for flights with either origin or destination as New York. The sudden increase in security can be the reason. Then there was a slight delay in months of November and December 2001 after the TSA was passed. The changes in procedures at airport can be the reason for that delay. After that time period delay decreased and there were no significant changes in it.

The regression analysis used to see the behavior of yearly trend across airlines support our initial hypothesis set there is a trend in the airlines delay from the year 2013 to 2016. There is a gradual increase in the time delay of airlines (Alaska Airline for which I have run the regression) across the years. This delay can be contributed to various reasons which include carrier delay, weather delay, national aviation system which controls the air traffic and may be because of tight security check delay. But if the trend of the delay across years for a airline under consideration shows a constant change in the time delay, it may be in the positive direction or in the negative direction. Both this trend can be highlighted to the airline management to look for if they improvement in their policy and system if there is continuous inclined nature in the time delay behavior.

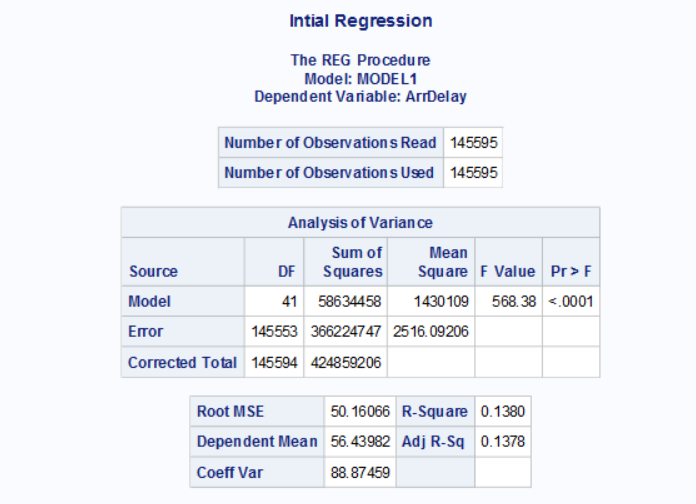
Our model confirms our hypothesis that along with the known factors of that location of the airports, Distance of travel, arrival time, departure time and air traffic that affect the airline industry the relatively less used factors of airline company and the effect of harsh weather in winter also affect the delay of flights.

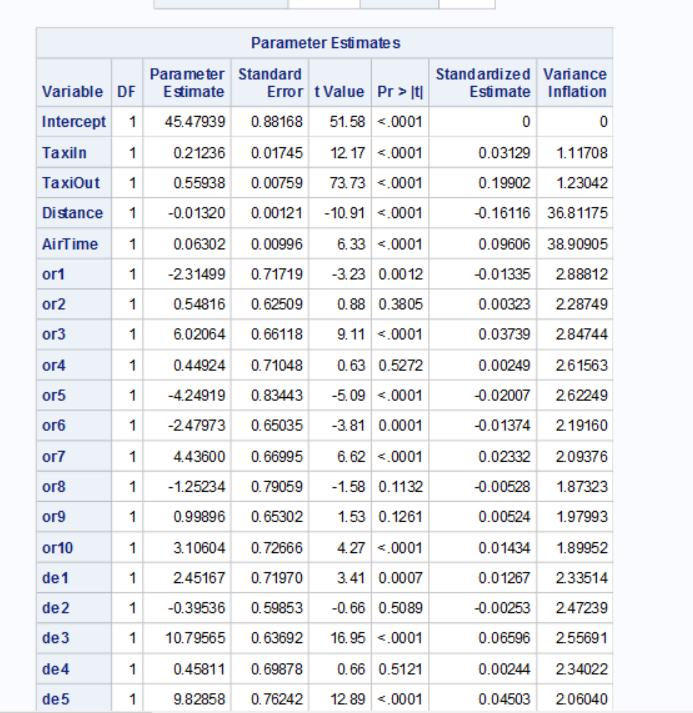
According to reports, Chicago Airport is notorious for delays. This can be interpreted from the model. Tuesdays and Wednesdays are better days for travelling. The period of midnight to early morning has the most delayed flights which may be to lower visibility. Winters would cause an increase in the average delay of flights as observed form the model. The model suggests that United Airlines and American Airlines were two of the worst performing airlines for the year 2008.

Despite these conclusions, we acknowledge that the results are not reliable due to specification bias and exclusion of many significant variables in the model. Also, we fail to consider the interaction between various categorical variables that could have an impact on the arrival delay.

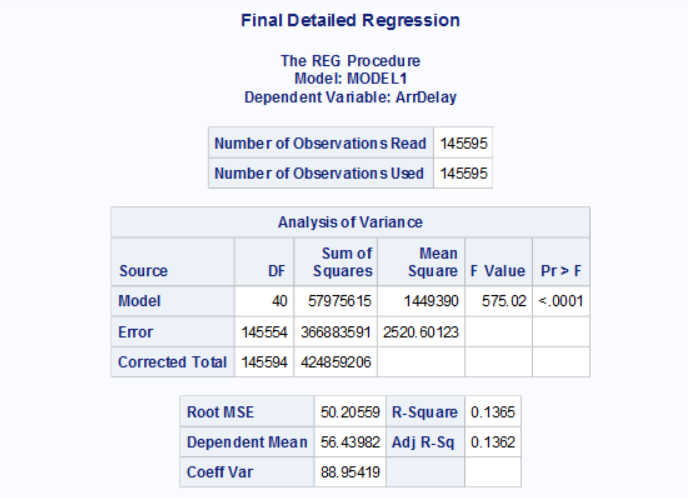
**Technical Appendix**

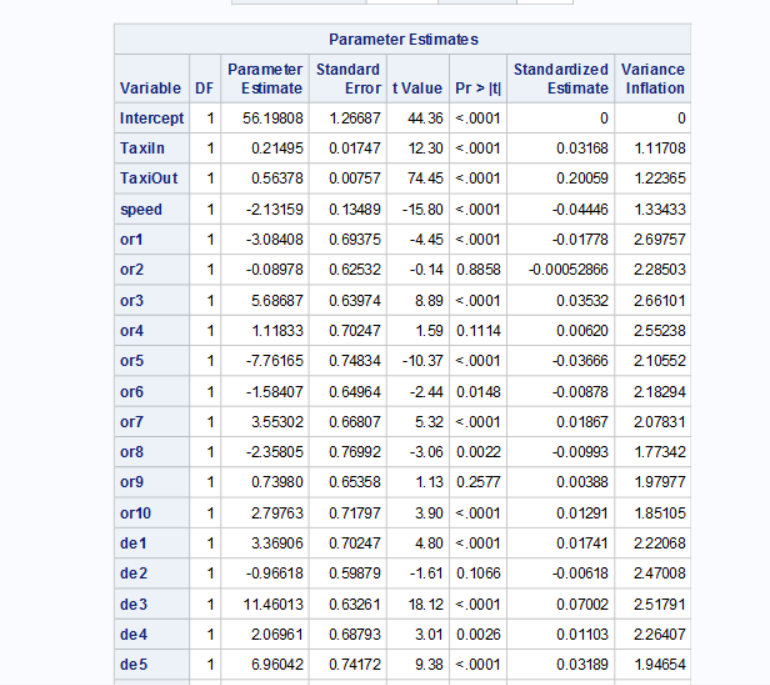
**Part-1**

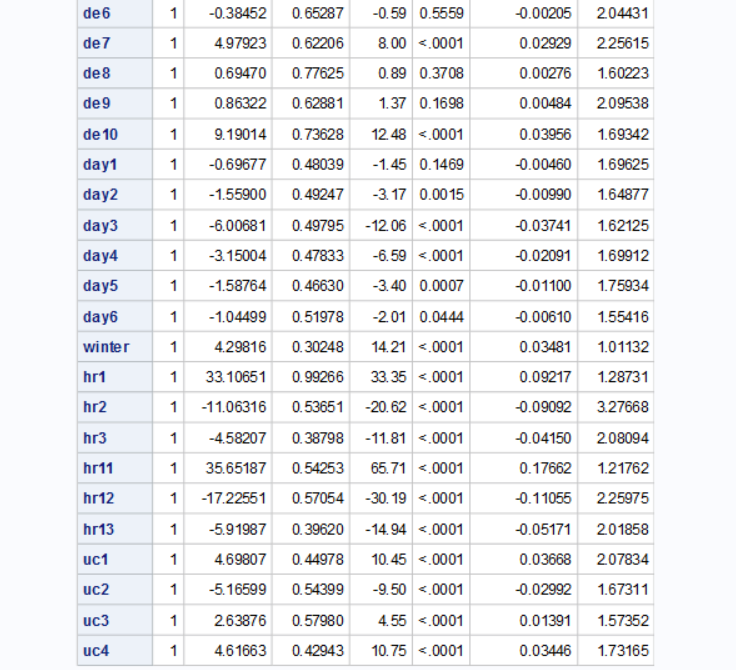


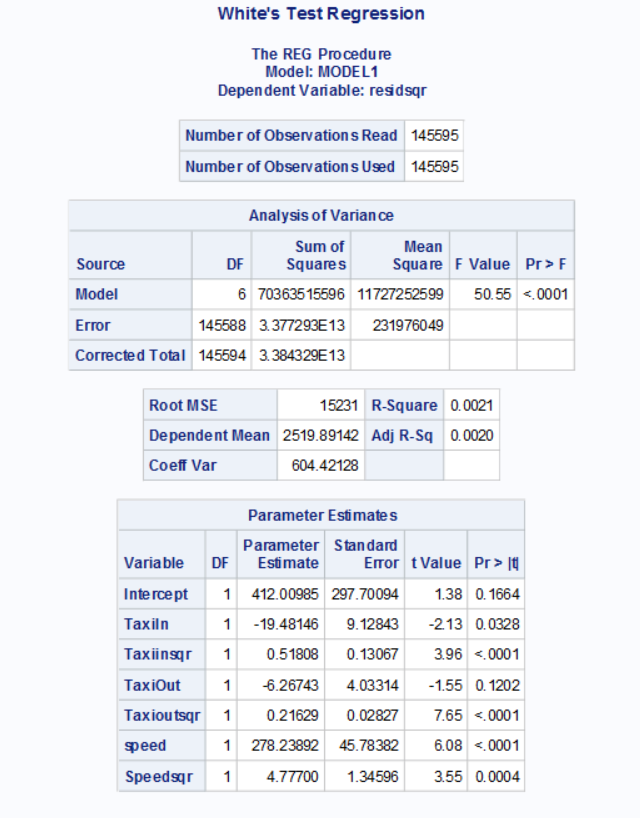












**Regression Equation**

**Definitions (**Include units of measurement here, e.g. thousands of dollars)

Y = arrdelay, in minutes

X2 = TaxiIn **time** The difference between the Wheels On time and Gate In time, in minutes.

X*3* = TaxiOut time spent by a flight between its actual off-block**time** (AOBT) and actual take-off **time** (ATOT), in minutes

X*4* = Speed in miles/min

D*1-* D*10* = Dummy variable for Origin of the airport

(‘ATL’ ‘LAX’ ‘ORD’ ‘DFW’ ‘JFK’ ‘DEN’ ‘SFO’ ‘CLT’ ‘LAS’ ‘IAH’ )

D*11-* D*20* = Dummy variable for Destination of the airport

(‘ATL’ ‘LAX’ ‘ORD’ ‘DFW’ ‘JFK’ ‘DEN’ ‘SFO’ ‘CLT’ ‘LAS’ ‘IAH’ )

D*21-* D*26* = Dummy for day of the week

D*27* = Dummy for Winter

D*28-* D*30*  = Dummies for the Arrival hours

(00:00 to 6:00, 6:00  to 12:00 , 12:00 to 18:00, 18:00 to 00:00)

D*31-* D33 = Dummies for the Departure hours

(00:00 to 6:00, 6:00  to 12:00 , 12:00 to 18:00, 18:00 to 00:00)

D*34-* D*37*  = Dummies for the airline company

**(F) Significance Test**

H0: β2 = β3 = … = β41 = 0 H1: β2 ≠ β3 ≠ … ≠ β41 ≠ 0

Given an F value of **575.02,** and a p-value of **less than**, we can reject H0 at the **0.05** level of significance, indicating that there is statistically significant evidence of a relationship between Y and the independent variables.

**Coefficient of Determination**

R2 =**13.65**. This value suggests that **13.65** percent of the variation in **arrdelay** is due to changes in the independent variables.

= **13.62**. This value suggests that **13.62** percent of the variation in **arrdelay** is due to changes in the independent variables, after adjusting for the number of independent variables.

**(T) Significance Tests**

H0: β2 = 0 H1: β2 ≠ 0

Given a t-value of **12.30** and a p-value of **0.0001,** we can reject H0 at the **0.05** level of significance, indicating that there is statistically significant evidence of a relationship between **arrdelay** and **TaxiIn** holding the effects of other independent variables constant. The standard error of β2 indicates the average error in estimating β2 is **0.01747**.

H0: β3 = 0 H1: β3 ≠ 0

Given a t-value of **74.45** and a p-value of **0.0001**, we can reject H0 at the **0.05** level of significance, indicating that there is statistically significant evidence of a relationship between **arrdelay** and **Taxiout** holding the effects of other independent variables constant. The standard error of β3 indicates the average error in estimating β3 is **0.00757**.

H0: β4 = 0 H1: β4 ≠ 0

Given a t-value of **-15.80** and a p-value of **less than 0.0001**, we can reject H0 at the **0.05** level of significance, indicating that there is statistically significant evidence of a relationship between **arrdelay** and **Taxiout** holding the effects of other independent variables constant. The standard error of β3 indicates the average error in estimating β3 is **0.13489**

H0: β7 = 0 H1: β7 ≠ 0

Given a t-value of **8.89** and a p-value of **less than** **0.0001**, we can reject H0 at the **0.05** level of significance, indicating that there is statistically significant evidence of a relationship between **arrdelay** and **or3(Chicago airport dummy)** holding the effects of other independent variables constant. The standard error of β7 indicates the average error in estimating β7 is **0.63974**.

H0: β13 = 0 H1: β13 ≠ 0

Given a t-value of **18.12** and a p-value of **less than 0.0001**, we can reject H0 at the **0.005** level of significance, indicating that there is statistically significant evidence of a relationship between **arrdelay** and **de3(Chicago airport dummy)** holding the effects of other independent variables constant. The standard error of β13 indicates the average error in estimating β13 is **0.63261.**

H0: β23 = 0 H1: β23 ≠ 0

Given a t-value of **-12.06** and a p-value of **less than 0.0001**, we can reject H0 at the **0.005** level of significance, indicating that there is statistically significant evidence of a relationship between **arrdelay** and **day3(Dummy for wednesday)**holding the effects of other independent variables constant. The standard error of β23 indicates the average error in estimating β23 is **0.49795**.

H0: β27 = 0 H1: β27 ≠ 0

Given a t-value of **-14.21** and a p-value of **less than 0.0001**, we can reject H0 at the **0.005** level of significance, indicating that there is statistically significant evidence of a relationship between **arrdelay** and **winter**holding the effects of other independent variables constant. The standard error of β27 indicates the average error in estimating β23 is **0.30248.**

H0: β28 = 0 H1: β28 ≠ 0

Given a t-value of **-33.10551** and a p-value of **less than 0.0001**, we can reject H0 at the **0.005** level of significance, indicating that there is statistically significant evidence of a relationship between **arrdelay** and **hr1** holding the effects of other independent variables constant. The standard error of βk indicates the average error in estimating **β28**is **33.10651**.

**Interpretation of Parameter Estimates**

= is the intercept where the regression plane crosses the Y-axis. However, because this represents the estimated value of **arrdelay** when all independent variables are equal to zero, this interpretation is reliable only when the data contains values of zero for each independent variable.

= A one minute increase in **TaxiIn** would result in a **0.21495 minutes**  change in **arrdelay**, holding the effects of other independent variables constant.

= A one minute increase in **TaxiIout** would result in a **0.56378 minutes** change in **arrdelay*,*** holding the effects of other independent variables constant.

= A one mile per minute increase in **speed** would result in a **-2.13159 minutes** change in **arrdelay*,*** holding the effects of other independent variables constant.

**Standardized Regression Coefficients (STBs)**

The standardized regression coefficients indicate that **TaxiOut** is contributing the greatest change/effect on our dependent variable

**Standard Error of the Regression (Root MSE)**

The standard error of the regression indicates that the average error in predicting Y using the regression equation is **50.20559**.

**Coefficient of Variation**

Our coefficient of variation compares the average error in predicting Y to the value of the dependent variable mean, and indicates that our model will have and average error of **88.895%**w hen predicting the value of our dependent variable.

**Plot Interpretation**

Not much can be read from the plot interpretation as the number of observations are too high to plot.

**Tests for Multicollinearity and Other Issues**

The initial regression had a case of multicollinearity. This could be deduced based on the following isuues:

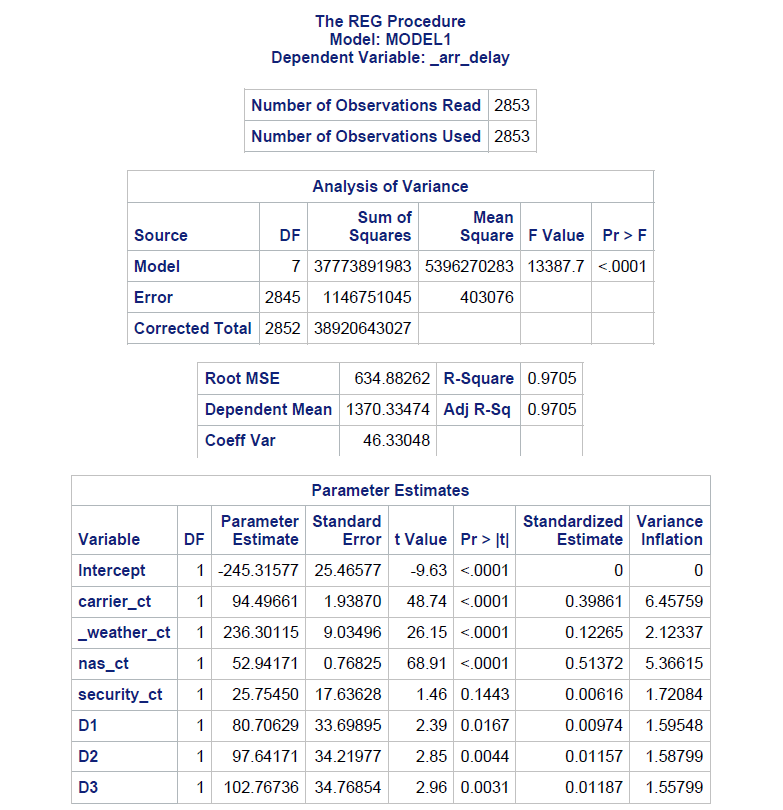
1. The Variance inflation of Distance and Airtime was high.
2. The condition index was above 10 indicating a severe case of near multicollinearity.
3. We transformed our variables by dividing Distance with Airtime to get the speed. This solved the problem of multicollinearity.

**Other Relevant Tests and Procedures**

Since the data is not a time series data we do not need to worry about autocorrelation. This can also be understood from the value of Durbin Watson of 1.808 which is very close to 2.

There is no heteroscedascity in data which can be confirmed by the white’s test where the calculated chi squared value is 305.74 which is much greater than the critical chi squared value of 11.07 at 0.05 level of confidence.

**Part-2**



**Regression Equation**

Y = β1+ β2X1+ β3X2+ β4X3+ β5X3+ β5X4 + β6D1+ β7D2 + β8 D3 + u

**Definitions**

Y =arrival delay

X1: CARRIER\_CT (percent of aircraft delayed due to carrier effect)

X2: WEATHER\_CT (percent of aircraft delayed due to weather effect)

X3: NAS\_CT (percent of aircraft delayed due National Aviation system effect)

X4: SECURITY\_CT (percent of aircraft delayed due to security effect)

D1: dummy variable for the year 2016

D2: dummy variable for the year 2015

D3: dummy variable for the year 2014

β1: intercept that take care of the value in 2013.

**(F) Significance Test**

H0: β2 = β3 = β4 = β5 = β6= β7= β8=0

H1: β2 ≠ β3 ≠ β4 ≠ β5 ≠ β6 ≠ β7 ≠ β8 ≠ 0

Given an F value of 13387.7, and a p-value of 0.001, we can reject H0 at the 0.05 level of significance, indicating that there is statistically significant evidence of a relationship between dep\_delay and independent variables.

**Coefficient of Determination R2** = 0.9705. This value suggests that 0.50 percent of the variation in dep\_delay is due to changes in independent variables.

**R̅2** = 0.9705. This value suggests that 0.55 percent of the variation in delay is due to changes in the independent variables, after adjusting for the number of independent variables.

**(T) Significance Tests**

H0: β2 =0

H1: β2 ≠ 0

Given a t-value of **48.74** and a p-value of **0.0001,** we can reject H0 at the **0.01** level of significance, indicating that there is statistically significant evidence of a relationship between **arr\_delay** and **carrier\_ct** holding the effects of other independent variables constant. The standard error of β2 indicates the average error in estimating β2 is **1.93870**

H0: β3 = 0

H1: β3 ≠ 0

Given a t-value of **26.15** and a p-value of **0.001,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **arr\_delay** and **weather\_ct** holding the effects of other independent variables constant. The standard error of β3 indicates the average error in estimating β3 is **9.034**

H0: β4 = 0

H1: β4 ≠ 0

Given a t-value of **68.91** and a p-value of **0.001,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **arr\_delay** and **nas\_ct** holding the effects of other independent variables constant. The standard error of β4 indicates the average error in estimating β4 is 0.76825

H0: β5 = 0

H1: β5 ≠ 0

Given a t-value of **1.46** and a p-value of **0.001,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **arr\_delay** and **security\_ct** holding the effects of other independent variables constant. The standard error of β5 indicates the average error in estimating β5 is **17.636**

H0: β6 = 0

H1: β6 ≠ 0

Given a t-value of **2.39** and a p-value of **0.0167,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **arr\_delay**and **D1** holding the effects of other independent variables constant. The standard error of β5 indicates the average error in estimating β5 is **33.6989**

H0: β7 = 0

H1: β7 ≠ 0

Given a t-value of **2.85** and a p-value of **0.0044,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **arr\_delay**and **D2** holding the effects of other independent variables constant. The standard error of β5 indicates the average error in estimating β5 is **0.0031**

H0: β8 = 0

H1: β8 ≠ 0

Given a t-value of **2.96** and a p-value of **0.0031,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **arr\_delay**and **D3** holding the effects of other independent variables constant. The standard error of β5 indicates the average error in estimating β5 is **34.76**

**Interpretation of Parameter Estimates**

= is the intercept where the regression plane crosses the Y-axis. However, because this represents the estimated value of **arr\_delay** when all independent variables are equal to zero, this interpretation is reliable only when the data contains values of zero for each independent variable.

= A one unit change in carrier\_ct would result in a **94.49** change in **arr\_delay** holding the effects of other independent variables constant.

= A one unit change in weather\_ct would result in a **236.30** change in **arr\_delay** holding the effects of other independent variables constant.

= A one unit change in nas\_ct would result in a **52.94** change in **arr\_delay** holding the effects of other independent variables constant.

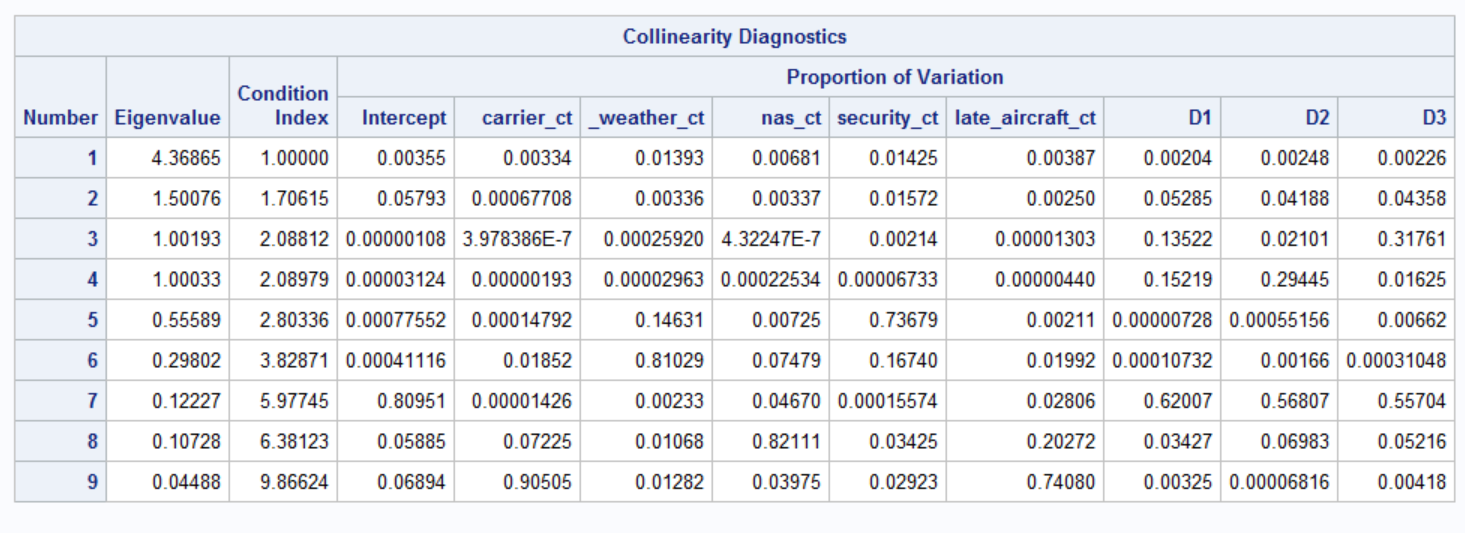
= A one unit change in security\_ct would result in a **25.75** change in **arr\_delay** holding the effects of other independent variables constant.

**Standard Error of the Regression (Root MSE)**  
The standard error of the regression indicates that the average error in predicting Y using the regression equation is 634.8826

**Coefficient of Variation**  
Our coefficient of variation compares the average error in predicting Y to the value of the dependent variable mean, and indicates that our model will have and average error of 46.33048 when predicting the value of our dependent variable.

**Test for Multicollinearity**

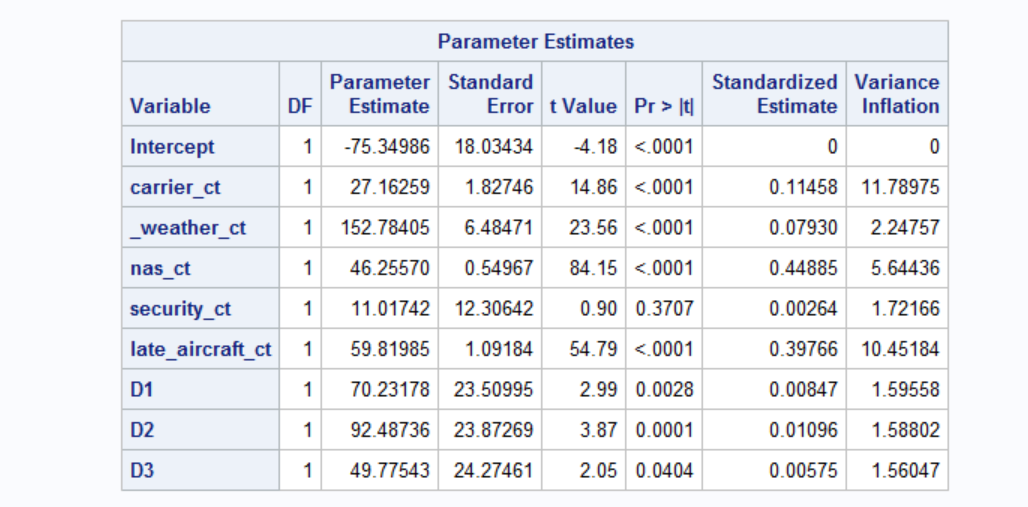
Output for Condition index



1. We have high pairwise correlation between carrier\_ct and late\_carrier\_ct (r= 0.96405) which is an indication of NMC.

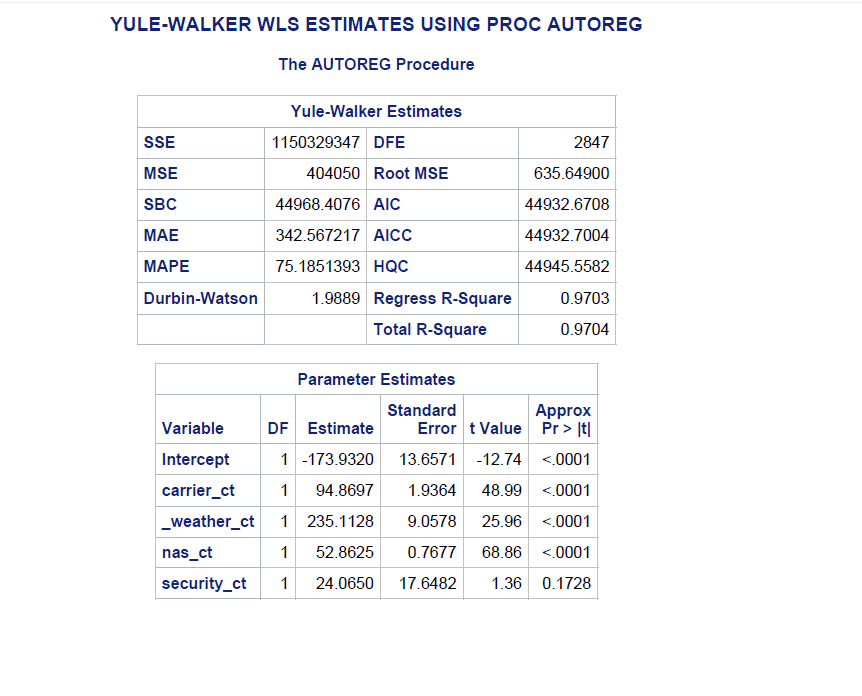
2. We have a High value of R2 and all the t-ratios are significantly low. So, this also indicates the problem of NMC.

1. The variance inflation of carrier\_ct, Dummy, and late\_carrier\_ct are also above 10, which indicates the presence of NMC.



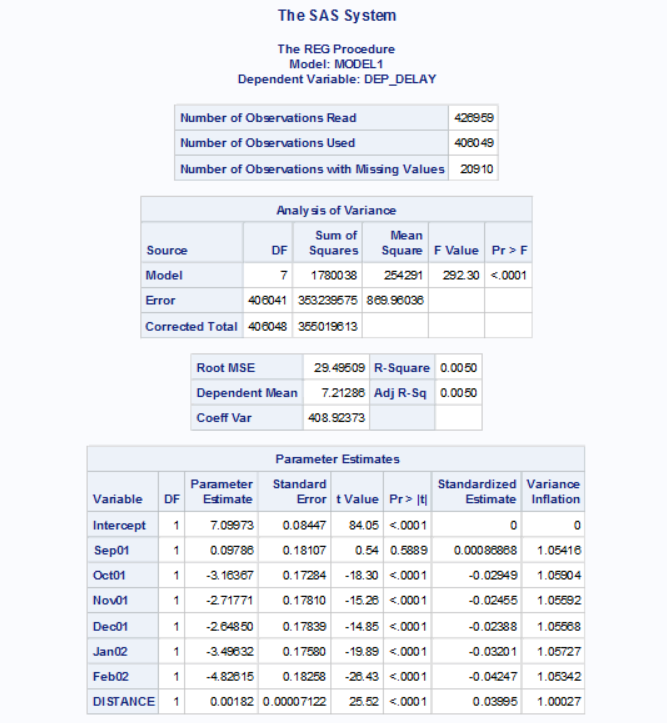
5. This condition index also indicates a case of severe near multicollinearity as the condition index is greater than 10 (condition index is 9.8862).

**AUTOCORRELATION TEST**



Since the Durbin-Watson value is near 2 (i.e. 1.9889). We can reject null hypothesis of auto-correlation.

**Part-3**



**Regression Equation**

Dep\_delay=  β1+ β2Sep01+ β3Oct01+ β4Nov01+ β5Dec01+ β6Jan02+ β7Feb02 + u

**Definitions**

Dep\_delay= departure delay

Sep01=Dummy variable for September 2001 month

Oct01=Dummy variable for October 2001 month

Nov01=Dummy variable for November 2001 month

Dec01=Dummy variable for December 2001 month

Jan02=Dummy variable for January 2002 month

Feb02=Dummy variable for February 2002 month

**(F) Significance Test**

H0: β2 = β3 = β4 = β5 = β6= β7= β8=0 H1: β2 ≠ β3 ≠ β4 ≠ β5 ≠ β6 ≠ β7 ≠ β8 ≠0

Given an F value of 292.30, and a p-value of 0.001, we can reject H0 at the 0.05 level of significance, indicating that there is statistically significant evidence of a relationship between dep\_delay and independent variables.

**Coefficient of Determination R2** = 0.0050. This value suggests that 0.50 percent of the variation in dep\_delay is due to changes in independent variables.

**R̅2** = 0.0055. This value suggests that 0.55 percent of the variation in delay is due to changes in the independent variables, after adjusting for the number of independent variables.

**(T) Significance Tests**

H0: β2 =0

H1: β2 ≠ 0

Given a t-value of **.54** and a p-value of **0.5889,** we can accept H0 at the **0.01** level of significance, indicating that there is no statistically significant evidence of a relationship between **dep\_delay** and **Sep01** holding the effects of other independent variables constant. The standard error of β2 indicates the average error in estimating β2 is **0.1810**

H0: β3 = 0

H1: β3 ≠ 0

Given a t-value of **-18.30** and a p-value of **0.001,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **dep\_delay** and **Oct01** holding the effects of other independent variables constant. The standard error of β3 indicates the average error in estimating β3 is **0.1728**

H0: β4 = 0

H1: β4 ≠ 0

Given a t-value of **-15.26** and a p-value of **0.001,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **dep\_delay** and **Nov01** holding the effects of other independent variables constant. The standard error of β4 indicates the average error in estimating β4 is 0.1781

H0: β5 = 0

H1: β5 ≠ 0

Given a t-value of **-14.8**5 and a p-value of **0.001,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **dep\_delay**and **Dec01** holding the effects of other independent variables constant. The standard error of β5 indicates the average error in estimating β5 is **0.1783**

H0: β6 = 0

H1: β6 ≠ 0

Given a t-value of **-19.89** and a p-value of **0.001,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **dep\_delay**and **Jan02** holding the effects of other independent variables constant. The standard error of β6 indicates the average error in estimating β6 is **0.1758**

H0: β7 = 0

H1: β7 ≠ 0

Given a t-value of **-4.8**2 and a p-value of **0.001,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **dep\_delay**and **Feb02** holding the effects of other independent variables constant. The standard error of β7 indicates the average error in estimating β7 is **0.1825**

H0: β8 = 0

H1: β8 ≠ 0

Given a t-value of **25.52** and a p-value of **0.001,** we can reject H0 at the **0.01** level of significance, indicating that there is a statistically significant evidence of a relationship between **dep\_delay** and **distance** holding the effects of other independent variables constant. The standard error of β8 indicates the average error in estimating β8 is **0.000017**

**Interpretation of Parameter Estimates**

= is the intercept where the regression plane crosses the Y-axis. However, because this represents the estimated value of **depdelay** when all independent variables are equal to zero, this interpretation is reliable only when the data contains values of zero for each independent variable.

= A one unit change in distance would result in a **0.0018** change in **depdelay** holding the effects of other independent variables constant.

Other estimates have not been mentioned because our analysis do not aim at prediction and the interpretation for our dummy variables does not make sense of what we are trying to prove.

**Standard Error of the Regression (Root MSE)**  
The standard error of the regression indicates that the average error in predicting Y using the regression equation is 29.49

**Coefficient of Variation**  
Our coefficient of variation compares the average error in predicting Y to the value of the dependent variable mean, and indicates that our model will have and average error of 408.92 when predicting the value of our dependent variable.