Statistical Computing Final Project Report

By: Nirupam Sharma
Batch: MS-BA Fall 2017
UCID: M12384101

SUMMARY

Aim of the project is to analyze and understand the key factors that affect the landing distance of an aircraft. The data sets used for analysis contained information regarding the aircraft type (data for Airbus and Boeing were used), speed in air, speed on ground, height, pitch, duration of flight and landing distance. Landing distance is the key variable to be predicted while other variables act as predictors. We used SAS programming language to perform the analysis. The analysis was completed in many phases. First phase was data loading and merging of tables together. Second phase was descriptive study of each variable followed by removal of duplicate values and abnormal data. The abnormal was stored separately which can be analyzed in future to infer certain patterns.

In third phase, descriptive analysis of data was carried out where we analyzed descriptive measures and histograms of various variables in the data. This analysis was further continued with hypothesis testing using t-test and ANOVA where we obtained a very critical information that landing distance varies a lot across the two aircrafts. After thorough analysis through plots between two variables and after performing Pearson correlation test we inferred that landing distance has very high correlation with speed in air and speed on ground.

In the last phase, Liner Regression models were built using both single and multiple variables. The selection of key variables of analysis were based on the findings from previous steps. We inferred from the models that speed on ground, speed in air, height and the aircraft type are the strong predictor variables. Two good models were obtained with high R-squared value explaining majority of variation in the data. The two models differed in a way that model contained Speed in air while the other contained speed on ground. We would prefer model that contains speed in air as it has more linear relationship with lading distance.

The final equation of the model is as follows.

Distance = -5954.02 + 13.79***Height*** + 80.2***Speed_air*** + 433.74***aircraft_value***

Project Details

BANA 6043 PROJECT

NAME: Nirupam Sharma **UCID:** M12384101

Background: Flight Landing

Goal: To study what factors and how they would impact the landing distance of a commercial flight.

Data: Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). See two Excel files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights).

Variable dictionary:

Aircraft: The make of an aircraft (Boeing or Airbus).
Duration (in minutes): Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min.
No_pasg: The number of passengers in a flight.
Speed_ground (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.
Speed_air (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.
Height (in meters): The height of an aircraft when it is passing over the threshold of the runway. The landing aircraft is required to be at least 6 meters high at the threshold of the runway.
Pitch (in degrees): Pitch angle of an aircraft when it is passing over the threshold of the runway.
Distance (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

Chapter 1: Data Preparation

Objective

Data preparation step consists of the following steps:

1. **Data collection and loading:** Find and obtain the data and load into SAS software environment
2. **Data merge:** Merge data from multiple tables if required using common attributes by combining tables vertically or horizontally.
3. **Data quality verification:** In this step, we check for missing and abnormal data across various columns and rows.
4. **Data cleaning:** After we identify missing and abnormal data, we either delete the missing and duplicate data or impute those missing value. As for abnormal data rows, we remove them from primary data and store them separately for further analysis to analyze presence of key information.
5. **Descriptive Summary of data:** We obtain the summary statistics of data such as number of observations, number of missing data points, mean, median, quartiles, standard deviation, variance and similar parameters.

SAS code

```
/* Connecting to first excel file.*/
PROC IMPORT OUT=FAA1
DATAFILE="/folders/myfolders/GASUE34_data/FAA1.xls"
        DBMS=xls REPLACE;
         GETNAMES=YES;
RUN;
proc print data=FAA1(obs=4);
Title FAA1_Data;
run;
/* Connecting to second excel file.*/
PROC IMPORT OUT=FAA2
DATAFILE="/folders/myfolders/GASUE34_data/FAA2.xls"
        DBMS=xls REPLACE;
         GETNAMES=YES;
RUN;
proc print data = FAA2(obs=4);
Title FAA2_Data;
run;

/*
Sorted first data set
 */
PROC SORT data=FAA1 out=stfaa1 ;
BY aircraft;
proc print data=stfaa1;
run;
```

```
Sorted second data set
 */
PROC SORT data=FAA2 out=stfaa2;
BY aircraft;
proc print data=stfaa2;
run;
/*
 Since second data contains 50 empty, so we remove them.
 */

data stfaa2;
   set stfaa2;
   if (_n_ > 50);
run;

/* Step1. Combining the two datasets. */

PROC SORT data=stfaa1;
BY aircraft; /*sorts the first data set. */
PROC SORT data=stfaa2;
BY aircraft; /*sorts the second data set. */
DATA combined_data;
SET stfaa1 stfaa2;
BY aircraft;
run;


Looking at summary of data
 */
proc means data=combined_data;
run;
/*
Removing duplicate values if any present
 */
PROC SORT data=combined_data NODUPKEY;
BY speed_ground speed_air height pitch distance;
RUN;
proc means data=combined_data;
title summary statistics before data cleaning
run;
/* Checking the number of missing values for every column. */
proc means data=combined_data n nmiss;
title Missing rows;
run;
/* The duration column contains 150 missing values
and speed_air contains 700 missing rows. We would look at the histogram
of both variables */
proc chart data=combined_data;
vbar speed_air duration;
run;
/*
The histogram of speed_air is right skewed while
that of duration is almost normal. Still I would
keep all the rows containing the missing values.
 */
```

```
/*
The histogram of speed_air is right skewed while
that of duration is almost normal. Still I would
keep all the rows containing the missing values.
checking abnormal values in the data by checking for
following rules.
1. Distance < 6000
2. Height > 6 meters
3. Speed_ground between 30mph and 140mph
4. Duration > 40 mins
I create a variable abnormal to indicate if any flight was abnormal or not
data validatephase1;
    set combined_data;
    if duration < 40 then abnormal = "yes";
    else if height < 6 then abnormal ="yes";
    else if speed_ground < 30 or speed_ground > 140 then abnormal="yes";
    else if distance > 6000 then abnormal ="yes";
    else abnormal ="no";
run;
/* checking the number of abnormal rows. */
PROC FREQ DATA=validatephase1;
TABLE abnormal;
Title Abnormal values before removal of abnormal entries;
RUN;
/* Since abnormal data can be useful for research so we
put the rows containing abnormal data separately. */
data abnormal_data;
    set validatephase1;
    if abnormal = "yes";
run;
PROC FREQ DATA=abnormal_data;
TABLE abnormal;
Title abnormal data;
RUN;
```

```
data complete_data;
    set validatephase1;
    if abnormal = "no";
run;
PROC FREQ DATA=complete_data;
TABLE abnormal;
Title Final prepared data;
RUN;
/* Removing temporary column abnormal. */
data final_data;
    set complete_data;
    Drop abnormal;
run;
proc means data=final_data;
title summary after data cleaning;
run;
/* Summarizing the data for different airlines.
1. Using proc univariate and plotting.*/
PROC SORT data=final_data;
BY aircraft;
proc univariate data=final_data normal plot;
    by aircraft;
    histogram ;
run;
/*2. Using proc means.
*/
proc means data= final_data n nmiss max min mean
                median mode q1 q3 std var;
    by aircraft;
run;
```

SAS output

Results: Project 1.sas

FAA1_Data

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584

FAA2_Data

Obs	aircraft	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	boeing	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584

Missing rows

The MEANS Procedure

Variable	Label	N	N Miss
duration	duration	800	50
no_pasg	no_pasg	850	0
speed_ground	speed_ground	850	0
speed_air	speed_air	208	642
height	height	850	0
pitch	pitch	850	0
distance	distance	850	0

Abnormal values before removal of abnormal entries

The FREQ Procedure

abnormal	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	781	97.63	781	97.63
yes	19	2.38	800	100.00

Abnormal values after removal of abnormal entries

The FREQ Procedure

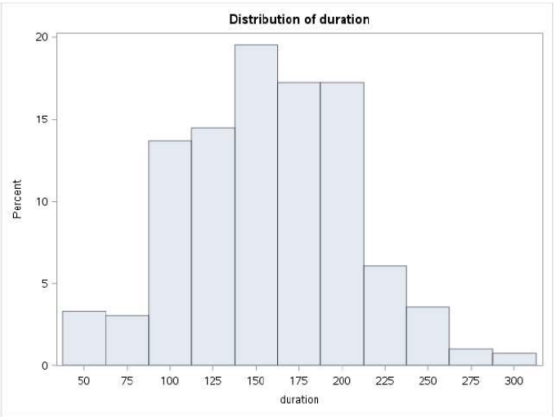
abnormal	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	781	100.00	781	100.00

The UNIVARIATE Procedure
Variable: duration (duration)

aircraft=airbus

Moments			
N	394	Sum Weights	394
Mean	156.90333	Sum Observations	61819.912
Std Deviation	49.1882899	Variance	2419.48786
Skewness	0.12708086	Kurtosis	-0.1808912
Uncorrected SS	10650608.8	Corrected SS	950858.73
Coeff Variation	31.3494238	Std Error Mean	2.47807024

Basic Statistical Measures			
Location		Variability	
Mean	156.9033	Std Deviation	49.18829
Median	156.4468	Variance	2419
Mode	.	Range	263.47548
		Interquartile Range	71.84843



Observations

- The distributions for almost all the variables are normal except for speed_air which has a right skewed distribution.
- The right skewness of the variable can be because it has many missing values.
- Although I deleted the duplicate rows, I think we need to have a unique identifier such as a primary key to differentiate the truly duplicate and unique rows.
- During the data preparation, I decided to leave the missing rows as they are because deleting them might lead to significant loss in information and it is always better to have bigger sample for modelling.
- Boeing seems to have on average higher values for most of the variables than the airbus.

Conclusions

We can conclude that the original data contained lot of missing data for especially one column speed_air and some 200 hundred abnormal rows which can be studied. Also, the distribution of duration is normal.

Chapter 2: Descriptive Study

Objective

The aim of chapter 2 is to understand the data and the relationships among the variables. We perform the following tasks in this step.

1. We perform statistical tests to check for certain hypothesis. The tests include t-test and anova.
2. We plot X-Y curves among the variables which help us understand relationships among the variables and this information can then we used in modelling step as well.
3. We find the correlation between variables. This can be useful in many ways as it can help us impute missing value of one variable from another and also for model optimization by removing one of the correlated variables.

SAS code

```
/*we check summary statistics of data*/
proc means data=final_data n nmiss mean median std var q1 q3 min max range;
title Summary Statistics;
/* we check for number of observations for each aircraft */
proc freq data=final_data;
table aircraft;
title number of observations in each aircraft;
/* checking the summary statistics for each aircraft separately */
proc means data= final_data n nmiss max min mean std var range;
    by aircraft;
title summary for each aircraft;
/* We now use t-test and ANOVA as data is balanced across aircraft carriers
is to check if the mean value of duration varies across aircrafts */
proc ttest data=final_data;
    class aircraft;
    var duration;
title t-test for aircraft and duration;
```

```
proc anova data=final_data;
    class aircraft;
    model duration = aircraft;
    means aircraft;
title ANOVA test for aircraft and duration;
/* we observe that the results of t-test and ANOVA
indicate that the means are similar and aircraft has no effect on duration.
We now check for correlation of duration with other numeric variables
and also among other variables as well*/
proc plot data = final_data;
 plot duration*no_pasg;
 plot duration*speed_ground;
 plot duration*speed_air;
 plot duration*height;
 plot duration*pitch;
 plot duration*distance;
 plot height*distance;
 plot speed_ground*speed_air;
/* from the X-Y plots we can see that duration has no clear
relation with other variables. Most of the graphs have spread data
indicating no clear relationship.
But linear relationship exists between speed_ground and speed_air.
we check for correlation between variables now*/
proc corr data= final_data;
    var duration no_pasg speed_ground speed_air height
    pitch distance;
title correlation among variables;
/* It can be concluded that speed_air, speed_ground and distance have
very high correlation among one another.*/
```

SAS output

Summary Statistics

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Median	Std Dev	Variance	Lower Quartile	Upper Quartile	Minimum	Maximum	Range
duration	duration	781	0	154.7757191	154.2845505	48.3499237	2337.72	119.6314577	189.6629425	41.9493694	305.6217107	263.6723414
no_pasg	no_pasg	781	0	60.0819462	60.0000000	7.5262579	56.6445583	55.0000000	65.0000000	29.0000000	87.0000000	58.0000000
speed_ground	speed_ground	781	0	79.6397499	79.7939604	18.8971690	357.1029943	66.1925304	92.1314349	33.5741041	132.7846766	99.2105726
speed_air	speed_air	195	586	103.5047686	100.8916770	9.8803757	97.6218233	96.1269654	109.4581269	90.0028586	132.9114649	42.9086063
height	height	781	0	30.4549525	30.2165682	9.7396415	94.8606171	23.5944766	36.9879836	6.2275178	59.9459639	53.7184462
pitch	pitch	781	0	4.0141289	4.0140064	0.5223688	0.2728692	3.6532968	4.3822934	2.2844801	5.9267842	3.6423041
distance	distance	781	0	1541.20	1273.66	904.5903306	818283.67	919.0474790	1960.43	41.7223127	5381.96	5340.24

number of observations in each aircraft

The FREQ Procedure

aircraft				
aircraft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
airbus	394	50.45	394	50.45
boeing	387	49.55	781	100.00

summary for each aircraft

The MEANS Procedure

aircraft=airbus

Variable	Label	N	N Miss	Maximum	Minimum	Mean	Std Dev	Variance	Range
duration	duration	394	0	305.6217107	42.1462262	156.9033299	49.1882899	2419.49	263.4754846
no_pasg	no_pasg	394	0	87.0000000	36.0000000	60.2868020	7.4864749	56.0473063	51.0000000
speed_ground	speed_ground	394	0	131.0351822	33.5741041	80.5319950	17.0601822	291.0498170	97.4610782
speed_air	speed_air	77	317	131.3379485	95.0113646	104.4454218	8.3303288	69.3943786	36.3265839
height	height	394	0	58.2277997	6.2275178	30.6001088	9.7740337	95.5317356	52.0002820
pitch	pitch	394	0	5.0373832	2.2844801	3.8268322	0.4856693	0.2358747	2.7529031
distance	distance	394	0	4896.29	41.7223127	1335.15	802.8381488	644549.09	4854.57

aircraft=boeing

Variable	Label	N	N Miss	Maximum	Minimum	Mean	Std Dev	Variance	Range
duration	duration	387	0	298.5223339	41.9493694	152.6096243	47.4467179	2251.19	256.5729646
no_pasg	no_pasg	387	0	82.0000000	29.0000000	59.8733850	7.5705312	57.3129427	53.0000000
speed_ground	speed_ground	387	0	132.7846766	33.8229533	78.7313660	20.5824990	423.6392648	98.9617233
speed_air	speed_air	118	269	132.9114649	90.0028586	102.8909526	10.7624189	115.8296605	42.9086063
height	height	387	0	59.9459639	7.5824946	30.3071707	9.7149204	94.3796779	52.3634694
pitch	pitch	387	0	5.9267842	2.9931514	4.2048134	0.4888554	0.2389796	2.9336328
distance	distance	387	0	5381.96	573.6217861	1750.98	953.8500300	909829.88	4808.34

t-test for aircraft and distance

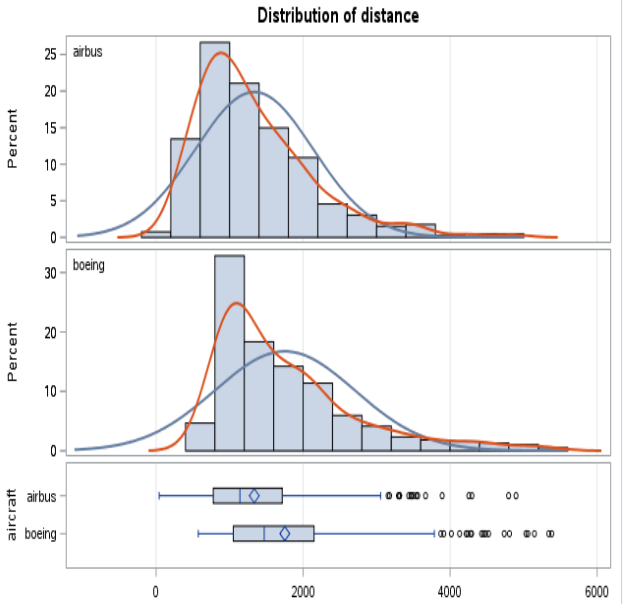
The TTEST Procedure
Variable: distance (distance)

aircraft	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus	394	1335.2	802.8	40.4464	41.7223	4896.3
boeing	387	1751.0	953.9	48.4869	573.6	5382.0
Diff (1-2)		-415.8	880.9	63.0452		

aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		1335.2	1255.6	1414.7	802.8	750.4	863.2
boeing		1751.0	1655.7	1846.3	953.9	891.1	1026.2
Diff (1-2)	Pooled	-415.8	-539.6	-292.1	880.9	839.3	926.9
Diff (1-2)	Satterthwaite	-415.8	-539.8	-291.9			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	779	-6.60	<.0001
Satterthwaite	Unequal	752.31	-6.59	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	386	393	1.41	0.0007



ANOVA test for aircraft and distance

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
aircraft	2	airbus boeing

Number of Observations Read	781
Number of Observations Used	781

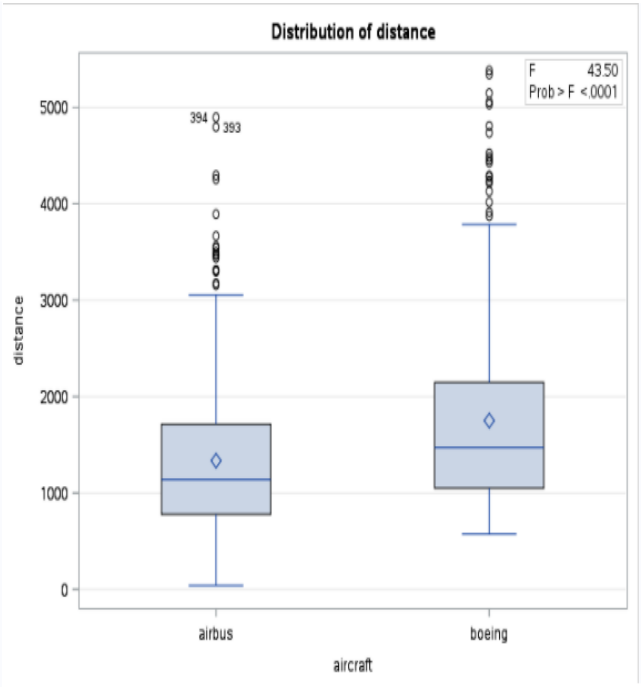
ANOVA test for aircraft and distance

The ANOVA Procedure
Dependent Variable: distance distance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	33759132.4	33759132.4	43.50	<.0001
Error	779	604502127.2	775997.6		
Corrected Total	780	638261259.6			

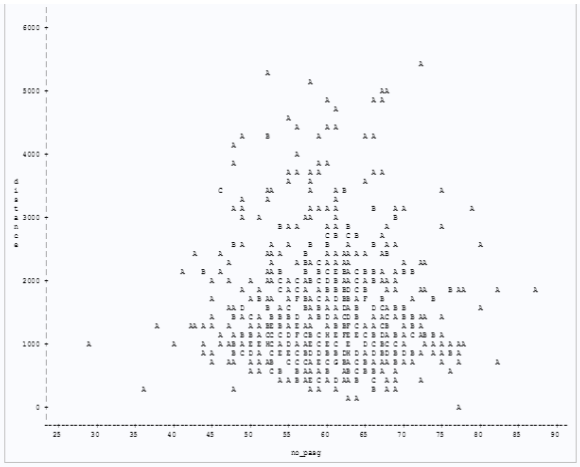
R-Square	Coeff Var	Root MSE	distance Mean
0.052892	57.15709	880.9073	1541.204

Source	DF	Anova SS	Mean Square	F Value	Pr > F
aircraft	1	33759132.39	33759132.39	43.50	<.0001

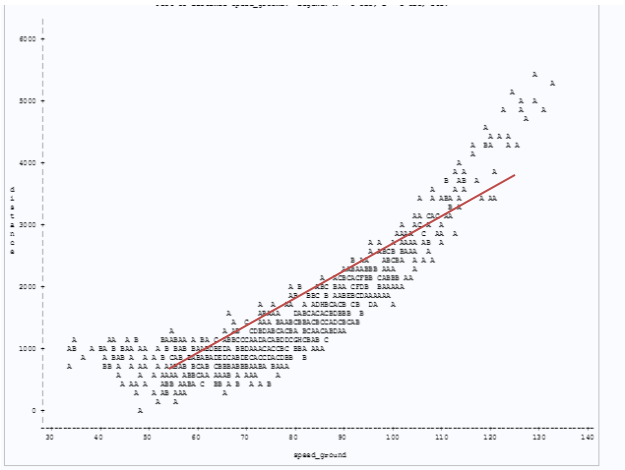


PLOTS

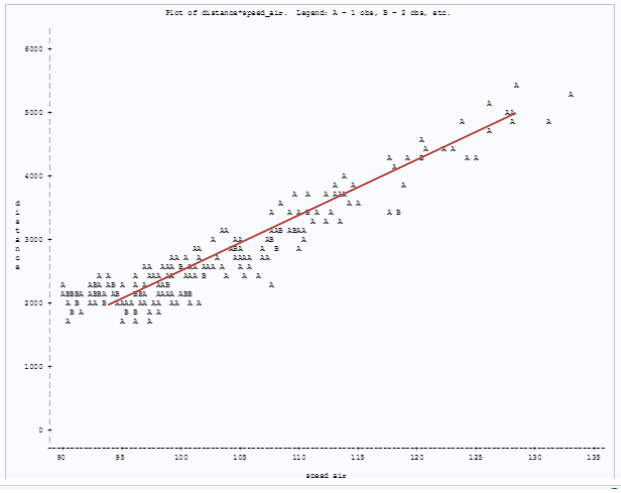
Distance vs no of passengers



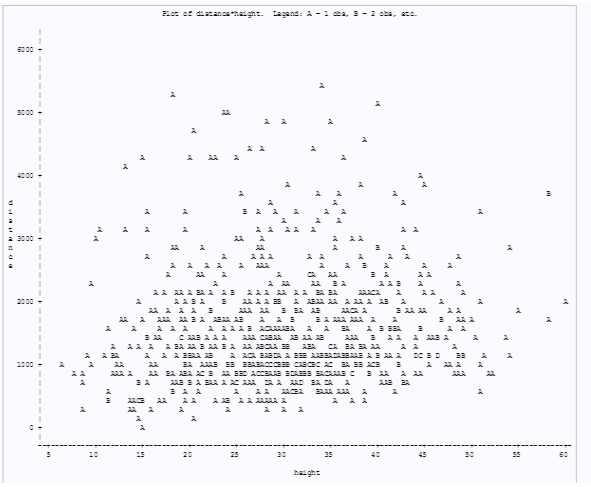
Distance vs speed_ground



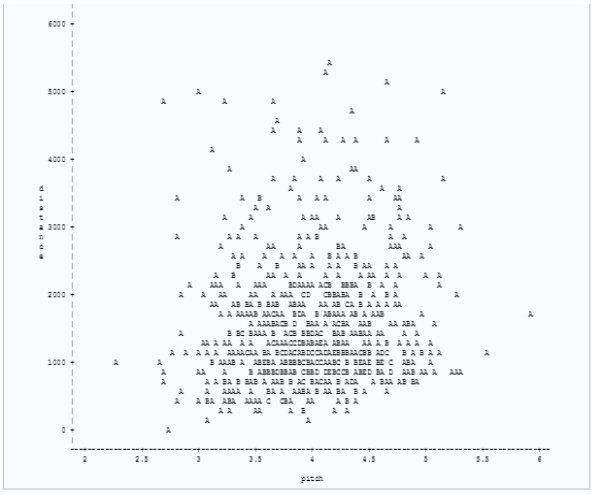
Distance vs speed_air



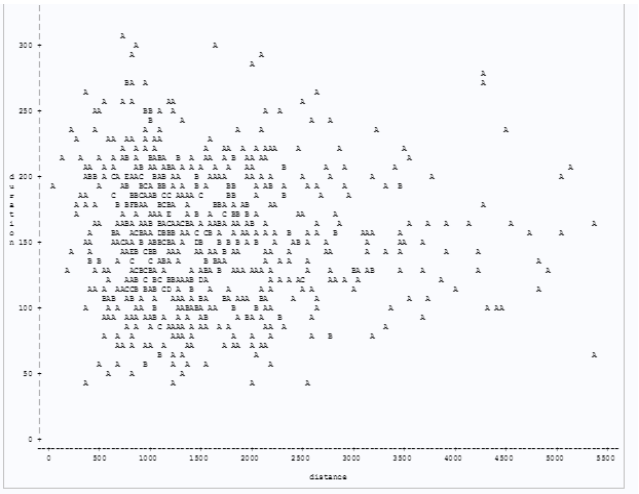
Distance vs Height



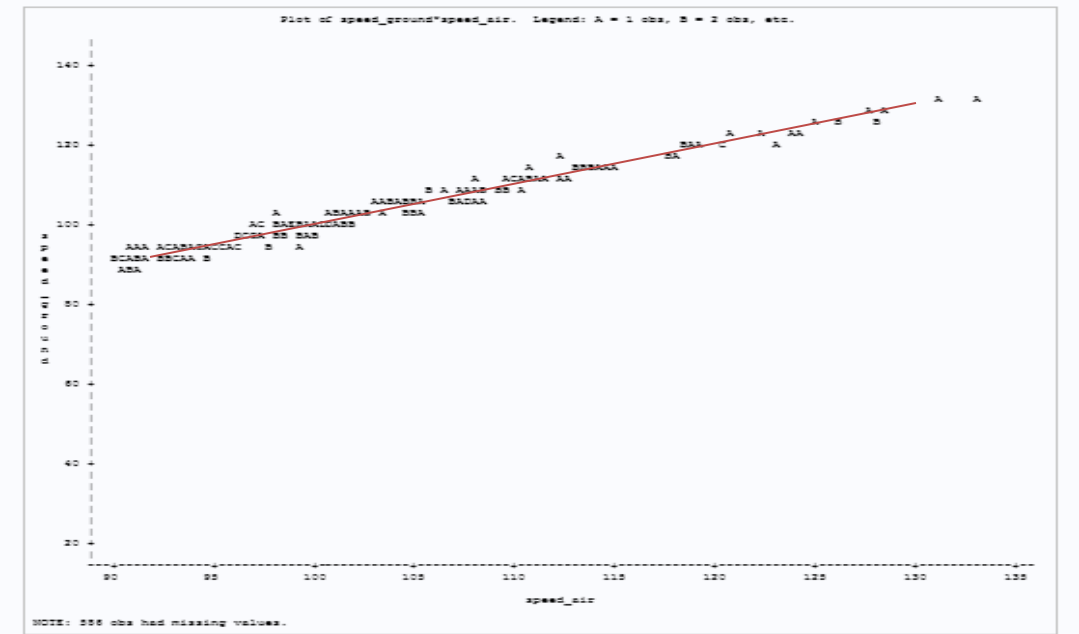
Distance vs Pitch



Duration vs Distance



Speed_ground vs Speed_air



Correlation Among Variables

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

	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
duration duration	1.00000 781	-0.03639 0.3098 781	-0.04897 0.1716 781	0.04454 0.5364 195	0.01112 0.7564 781	-0.04675 0.1918 781	-0.05138 0.1514 781
no_pasg no_pasg	-0.03639 0.3098 781	1.00000 781	-0.00149 0.9669 781	0.00002 0.9998 195	0.03731 0.2977 781	-0.01445 0.6869 781	-0.01685 0.6382 781
speed_ground speed_ground	-0.04897 0.1716 781	-0.00149 0.9669 781	1.00000 781	0.98835 <.0001 195	-0.05167 0.1491 781	-0.05167 0.1491 781	0.86771 <.0001 781
speed_air speed_air	0.04454 0.5364 195	0.00002 0.9998 195	0.98835 <.0001 195	1.00000 195	-0.08673 0.2280 195	-0.04827 0.5028 195	0.94322 <.0001 195
height height	0.01112 0.7564 781	0.03731 0.2977 781	-0.05167 0.1491 781	-0.08673 0.2280 195	1.00000 781	0.03474 0.3323 781	0.10372 0.0037 781
pitch pitch	-0.04675 0.1918 781	-0.01445 0.6869 781	-0.05167 0.1491 781	-0.04827 0.5028 195	0.03474 0.3323 781	1.00000 781	0.06868 0.0550 781
distance distance	-0.05138 0.1514 781	-0.01685 0.6382 781	0.86771 <.0001 781	0.94322 <.0001 195	0.10372 0.0037 781	0.06868 0.0550 781	1.00000 781

High Correlation

Observations

After performing descriptive analysis, the following observations were obtained:

1. There results of t-test and ANOVA indicate that aircraft has high effect on the landing distance.
2. After looking at the plots we can conclude there is strong relationships between following pairs of variables:
    a. Distance - Speed_air (correlation as .94322)
    b. Distance – Speed_ground (correlation as .86771)
    c. Speed_air – Speed_ground (correlation as .98835)

Conclusions

We can conclude that aircraft has a relation with landing distance. It can also be concluded that landing distance, speed on ground and speed are highly correlated.

This information will play a very critical role during model building phase.

Chapter 3: Statistical modelling

Objective

Modelling is the area where we try to represent the data in the form a mathematical equation. We try we form a model which can explain as much possible variation in data. Once a model is obtained, it can be applied in future to predict numeric and qualitative values. When the model is used to predict numerical data, the modelling technique comes under regression while when we try to solve to predict categorical variable then the technique is called classification.

Here the variable we are trying to predict is landing distance which is numerical hence we apply linear regression technique here.

We start with single variable linear regression and follow it by multi variable regression analysis here.

SAS code

```
/* We start with regression. For Regression I will create two separate data sets
also one for each aircraft.I will create models for each of these 3 data sets.
We would preferably use those variables which have high correlation with distance.*/
data boeingdata;
    set final_data;
    if aircraft = "boeing";
run;
data airbusdata;
    set final_data;
    if aircraft = "airbus";
run;

data boeingdata;
    set final_data;
    if aircraft = "boeing";
run;
data airbusdata;
    set final_data;
    if aircraft = "airbus";
run;
```

```
/* Model 1: Start with linear Regression using speed_ground as
predictor for whole data*/
proc reg data=final_data;
 model distance= speed_ground;
run; /* Return R-squared value of .7529 and adj-rsqr .7526

/* Model 2:linear Regression for whole data using speed_air*/
proc reg data=final_data;
 model distance= speed_air;
run;/* Return R-squared value of .8897 and adj-rsqr .8891

/* Model 3:linear Regression for whole data using aircraft_value*/
proc reg data=final_data;
 model distance= aircraft_value;
run;/* Return R-squared value of .0127 and adj-rsqr .0122

/* Model 4:linear Regression for whole data using height*/
proc reg data=final_data;
 model distance= height;
run;/* Return R-squared value of .0527 and adj-rsqr .0517

/* Model 5:Start with multivariate linear Regression using speed_ground and speed_air
for whole data*/
proc reg data=final_data;
 model distance= speed_ground speed_air;
run;/* Return R-squared value of .8902 and adj-rsqr .8802

/* Model 6:Start with multivariate linear Regression using speed_air and height
for whole data*/
proc reg data=final_data;
 model distance= height speed_air;
run;/* Return R-squared value of .9093 and adj-rsqr .9083
```

```sas
/* Model 7:Start with multivariate linear Regression using speed_ground and height
for whole data*/
proc reg data=final_data;
 model distance= height speed_ground;
run;/* Return R-squared value of .7707 and adj-rsqr .7704

/* Model 8:Multivariate linear Regression using aircraft_value, height
amd speed_air for whole data*/
proc reg data=final_data;
 model distance= height aircraft_value speed_air;
run;/* Return R-squared value of .9502 and adj-rsqr .9496

/* Model 9:Multivariate linear Regression using speed_ground, aircraft
and speed_air for whole data*/
proc reg data=final_data;
 model distance= speed_ground height speed_air aircraft_value;
run;/* Return R-squared value of .9505 and adj-rsqr .9497

/* Model 10:Multivariate linear Regression using all variables for whole data*/
proc reg data=final_data;
 model distance= speed_ground speed_air aircraft_value duration height pitch;
run;/* Return R-squared value of .9744 and adj-rsqr .9736

/* Model 11:Start with linear Regression using speed_ground as
predictor for boeingdata data*/
proc reg data=boeingdata;
 model distance= speed_ground;
run; /* Return R-squared value of .8109 and adj-rsqr .8104

/* Model 12:linear Regression for boeingdata data using speed_air*/
proc reg data=boeingdata;
 model distance= speed_air;
run;/* Return R-squared value of.9557 and adj-rsqr as .9553
```

```
/* Model 13:Start with multivariate linear Regression using speed_ground and speed_air
for boeingdata data*/
proc reg data=boeingdata;
 model distance= speed_ground speed_air;
run;/* Return R-squared value of .9557 and adj-rsqr .9549

/* Model 14:Start with multivariate linear Regression using speed_ground and height
for boeingdata data*/
proc reg data=boeingdata;
 model distance= speed_ground height;
run;/* Return R-squared value of .8317 and adj-rsqr .8308

/* Model 15:Start with linear Regression using speed_ground as
predictor for airbusdata data*/
proc reg data=airbusdata;
 model distance= speed_ground;
run; /* Return R-squared value of .8258 and adj r-sqr as .8254

/* Model 16:linear Regression for airbusdata data using speed_air*/
proc reg data=airbusdata;
 model distance= speed_air;
run;/* Return R-squared value of .9317 and adj r-sqr as .9309

/* Model 17:linear Regression for airbusdata data using height*/
proc reg data=airbusdata;
 model distance= height;
run;/* Return R-squared value of .0311 and adj r-sqr as .0302

/* Model 18:Start with multivariate linear Regression using speed_ground and speed_air
for airbusdata data*/
proc reg data=airbusdata;
 model distance= speed_ground speed_air;
run;/* Return R-squared value of .9341 and adj-rsqr as .9323
```

SAS output

Model 1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	480561690	480561690	2373.87	<.0001
Error	779	157699570	202438		
Corrected Total	780	638261260			

Root MSE	449.93163	R-Square	0.7529
Dependent Mean	1541.20394	Adj R-Sq	0.7526
Coeff Var	29.19352		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1766.75731	69.77693	-25.32	<.0001
speed_ground	speed_ground	1	41.53656	0.85252	48.72	<.0001

Model 2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	118926290	118926290	1556.17	<.0001
Error	193	14749519	76422		
Corrected Total	194	133675809			

Root MSE	276.44598	R-Square	0.8897
Dependent Mean	2784.49158	Adj R-Sq	0.8891
Coeff Var	9.92806		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5417.60675	208.86035	-25.94	<.0001
speed_air	speed_air	1	79.24368	2.00880	39.45	<.0001

Model 3

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	33759132	33759132	43.50	<.0001
Error	779	604502127	775998		
Corrected Total	780	638261260			

Root MSE	880.90726	R-Square	0.0529
Dependent Mean	1541.20394	Adj R-Sq	0.0517
Coeff Var	57.15709		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1750.98330	44.77903	39.10	<.0001
aircraft_value		1	-415.83167	63.04521	-6.60	<.0001

Model 4

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	118995206	59497603	778.14	<.0001
Error	192	14680603	76461		
Corrected Total	194	133675809			

Root MSE	276.51668	R-Square	0.8902
Dependent Mean	2784.49158	Adj R-Sq	0.8890
Coeff Var	9.93060		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5425.48760	209.07862	-25.95	<.0001
speed_ground	speed_ground	1	-12.31906	12.97593	-0.95	0.3436
speed_air	speed_air	1	91.62992	13.20051	6.94	<.0001

Model 5

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	127014334	63507167	1830.43	<.0001
Error	192	6661474	34695		
Corrected Total	194	133675809			

Root MSE	186.26642	R-Square	0.9502
Dependent Mean	2784.49158	Adj R-Sq	0.9496
Coeff Var	6.68942		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5418.01601	140.72793	-38.50	<.0001
aircraft_value		1	-417.87608	27.36909	-15.27	<.0001
speed_air	speed_air	1	80.84183	1.35755	59.55	<.0001

Model 6

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	127053772	42351257	1221.54	<.0001
Error	191	6622037	34670		
Corrected Total	194	133675809			

Root MSE	186.19976	R-Square	0.9505
Dependent Mean	2784.49158	Adj R-Sq	0.9497
Coeff Var	6.68703		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5423.97851	140.78861	-38.53	<.0001
speed_ground	speed_ground	1	-9.32137	8.73989	-1.07	0.2875
speed_air	speed_air	1	90.21152	8.88940	10.15	<.0001
aircraft_value		1	-417.21945	27.36622	-15.25	<.0001

Model 7

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	130257907	21709651	1194.13	<.0001
Error	188	3417901	18180		
Corrected Total	194	133675809			

Root MSE	134.83444	R-Square	0.9744
Dependent Mean	2784.49158	Adj R-Sq	0.9736
Coeff Var	4.84234		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5915.50780	139.04730	-42.54	<.0001
speed_ground	speed_ground	1	-3.63956	6.43387	-0.57	0.5723
speed_air	speed_air	1	85.64372	6.54021	13.09	<.0001
aircraft_value		1	-439.40658	21.29791	-20.63	<.0001
duration	duration	1	0.14822	0.20402	0.73	0.4684
height	height	1	13.68209	1.04149	13.14	<.0001
pitch	pitch	1	-12.94161	18.65656	-0.69	0.4887

Model 8

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	87718791	87718791	2502.43	<.0001
Error	116	4066204	35053		
Corrected Total	117	91784994			

Root MSE	187.22575	R-Square	0.9557
Dependent Mean	2899.87705	Adj R-Sq	0.9553
Coeff Var	6.45633		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5378.03263	166.37305	-32.33	<.0001
speed_air	speed_air	1	80.45323	1.60828	50.02	<.0001

Model 9

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	87718965	43859483	1240.48	<.0001
Error	115	4066029	35357		
Corrected Total	117	91784994			

Root MSE	188.03398	R-Square	0.9557
Dependent Mean	2899.87705	Adj R-Sq	0.9549
Coeff Var	6.48421		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5378.19445	167.10715	-32.18	<.0001
speed_ground	speed_ground	1	-0.81520	11.60953	-0.07	0.9441
speed_air	speed_air	1	81.26917	11.73166	6.93	<.0001

Model 10

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	209192675	209192675	1858.85	<.0001
Error	392	44115119	112539		
Corrected Total	393	253307794			

Root MSE	335.46769	R-Square	0.8258
Dependent Mean	1335.15163	Adj R-Sq	0.8254
Coeff Var	25.12581		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2108.83736	81.64851	-25.83	<.0001
speed_ground	speed_ground	1	42.76547	0.99191	43.11	<.0001

Model 11

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	35324269	35324269	1023.71	<.0001
Error	75	2587965	34506		
Corrected Total	76	37912234			

Root MSE	185.75846	R-Square	0.9317
Dependent Mean	2607.66711	Adj R-Sq	0.9308
Coeff Var	7.12355		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5940.18615	267.99593	-22.17	<.0001
speed_air	speed_air	1	81.84038	2.55788	32.00	<.0001

Model 12

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	35412328	17706164	524.12	<.0001
Error	74	2499906	33783		
Corrected Total	76	37912234			

Root MSE	183.80019	R-Square	0.9341
Dependent Mean	2607.66711	Adj R-Sq	0.9323
Coeff Var	7.04845		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5977.14845	266.15716	-22.46	<.0001
speed_ground	speed_ground	1	-21.49061	13.31087	-1.61	0.1107
speed_air	speed_air	1	103.67917	13.76127	7.53	<.0001

Model 13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6866417	6866417	8.47	0.0037
Error	779	631394842	810520		
Corrected Total	780	638261260			

Root MSE	900.28867	R-Square	0.0108
Dependent Mean	1541.20394	Adj R-Sq	0.0095
Coeff Var	58.41464		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1247.82243	105.82022	11.79	<.0001
height	height	1	9.63329	3.30972	2.91	0.0037

Model 14

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	121549666	60774833	962.28	<.0001
Error	192	12126142	63157		
Corrected Total	194	133675809			

Root MSE	251.31055	R-Square	0.9093
Dependent Mean	2784.49158	Adj R-Sq	0.9083
Coeff Var	9.02537		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5901.02410	204.14873	-28.91	<.0001
height	height	1	12.43014	1.92866	6.44	<.0001
speed_air	speed_air	1	80.26829	1.83306	43.79	<.0001

Model 15

Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	781
Number of Observations Used	781

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	494685306	247342653	1340.28	<.0001
Error	778	143575953	184545		
Corrected Total	780	638261260			

Root MSE	429.58693	R-Square	0.7751
Dependent Mean	1541.20394	Adj R-Sq	0.7745
Coeff Var	27.87346		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2217.42835	84.21585	-26.33	<.0001
height	height	1	13.83449	1.58140	8.75	<.0001
speed_ground	speed_ground	1	41.90500	0.81506	51.41	<.0001

Model 16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	130232033	43410678	2407.66	<.0001
Error	191	3443775	18030		
Corrected Total	194	133675809			

Root MSE	134.27672	R-Square	0.9742
Dependent Mean	2784.49158	Adj R-Sq	0.9738
Coeff Var	4.82231		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5954.38353	109.10498	-54.57	<.0001
height	height	1	13.79126	1.03236	13.36	<.0001
aircraft_value		1	-433.74061	19.76568	-21.94	<.0001
speed_air	speed_air	1	82.03932	0.98273	83.48	<.0001

Model 17

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	292076444	146038222	948.59	<.0001
Error	384	59117890	153953		
Corrected Total	386	351194334			

Root MSE	392.36824	R-Square	0.8317
Dependent Mean	1750.98330	Adj R-Sq	0.8308
Coeff Var	22.40845		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2008.46764	104.75662	-19.17	<.0001
speed_ground	speed_ground	1	42.28538	0.97362	43.43	<.0001
height	height	1	14.19682	2.06276	6.88	<.0001

Model 18

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	35412328	17706164	524.12	<.0001
Error	74	2499906	33783		
Corrected Total	76	37912234			

Root MSE	183.80019	R-Square	0.9341
Dependent Mean	2607.66711	Adj R-Sq	0.9323
Coeff Var	7.04845		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5977.14845	266.15716	-22.46	<.0001
speed_ground	speed_ground	1	-21.49061	13.31087	-1.61	0.1107
speed_air	speed_air	1	103.67917	13.76127	7.53	<.0001

Observations

The observations are as follows:

- **Models on subsets of data**: Initially, I made 3 different sets of data: whole data and data specific to being and airbus respectively. I also made a dummy variable *aircraft*_value to explain which kind of aircraft (Boeing as 1 and Airbus as 0) it is. After creating models to explain landing distance on each of these subsets, it could be inferred that the coefficients, critical predictors and R-squared value does not change much. So, we would continue with models built over entire data containing observations for both the aircrafts.
- **Single variable linear regression**: After building single variable linear models using each of the variables speed_air, speed_ground, height, pitch, and duration, we observe that R-squared value comes out to be maximum for speed_air followed by speed_ground. This is due the high correlation these variables have with the output variable distance.
- **Multivariate linear regression**: When we switch to multivariate linear regression models, following observations were seen.
    - When all the variables are taken to make the model the R-squared value is coming as .9744 and RMSE of 134. Although the R-squared value is high but this will lead to overfitting and the model might not give proper estimate on future data due to lack of variability.
    - The output P-value for each variable from above model is as follows:

Variable Name	**P-value**
Pitch	.4887
Duration	.4684
Speed_air	< .001
Speed_ground	< .001
Height	< .001
Aircraft_value	< .001

It can be concluded from one variable and multivariate model (containing all variables) that duration and pitch have high P-value so we **can't reject** the hypothesis that $\mathbf{B_{duration}=0}$ and $\mathbf{B_{pitch}=0}$. We **can reject** the hypothesis $\mathbf{B_{speed\_air}=0}$, $\mathbf{B_{speed\_ground}=0}$, $\mathbf{B_{height}=0}$ and $\mathbf{B_{aircraft\_value}=0}$.

- We create two separate models, one that contains speed_air and other that contains speed_ground. Both these models contain aircraft_value and height as well.

MODEL 1

Distance = -5954.02 + 13.79***Height*** + 80.2***Speed_air*** + 433.74***aircraft_value***

R-Squared value= .9794, RMSE=130, Obs used= 124

MODEL 2

Distance = -2217.42 + 13.84***Height*** + 41.91***Speed_ground*** + 360.97***aircraft_value***

R-Squared value=.8481, RMSE = 340, Obs used=781

Both the above models have very high R-squared values. Model 1 is based on considering only 195 observations as speed_air has missing values for other rows. Model 2, on the other hand, although a lower R-squared value but is built on larger sample of entire 781 observations.

- **Model Selection**: We can choose model based on either Speed_air or Speed_ground as they both have good R-squared value. Speed_air has perfect linear relationship with landing distance while relationship between Speed_ground and landing distance is less linear. One very important point to consider here is that we are only interested in those cases where the landing distance is more than 6000. For such cases values for both Speed_air and Speed_ground are present. Hence, we would choose Model 1 containing Speed_air over Model 2 containing Speed_ground.

Conclusions

We found that variables Speed_ground, Speed_ground, aircraft_value and Height are key features and predictor variables while pitch and duration are not good predictors for landing distance.

From all the above observations, the following multivariate Linear Regression model will be used to predict the value of landing Distance. We would choose the following model containing Speed_air as it has high collinearity with distance. In case we have missing speed_air in test data then we may use model 2 that contains Speed_ground.

The final model to be used for predicting landing distance is

Distance = -5954.02 + 13.79***Height*** + 80.2***Speed_air*** + 433.74***aircraft_value***

**Note: if Speed_air data available is available for test or input data than this model is to be used over second one.*

We use the below alternative version only if Speed_air is not present.

Landing Distance = -2217.42 + 13.84*Height + 41.91*Speed_ground + 360.97*aircraft_value

**Note: This model will only be used when Speed_air data is not present in the test or input data.*