BANA 6043 PROJECT

Name: Madhava Chandra **UCID:**

Background: Flight landing.

Motivation: To reduce the risk of landing overrun.

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).

Summary:

The purpose of this study was to determine a safe landing distance. The data was obtained from FAA and contained about 850 unique values pertaining to flight landing. Few of the key variables involved were flight speeds in air and ground, height of the flight prior to landing and duration of the flight. The data was first analyzed for outliers and then correlations were examined. Upon then weeding out the nonsignificant variables, the landing distance was regressed over the key variables and it was concluded that the speed in air and ground explain the landing distance. Also two models have been built and depending on the user preference, a more complex mix of models(based on whether the air speed is captured) or a simpler model based only on ground speed can be used to predict the landing distance.

1. Data exploration and data cleaning

Goal: This module focuses on understanding the variables present.

Observations and Decisions: (SAS code and outputs to follow)

- First summary of the loaded data was observed for each of FAA1 and FAA2.
- It was then concluded that there are duplicates in this data along with empty observations.
- Post removing these, the proc means was studied for each of these to conclude that these values belong to the same population, based on which the data was combined.
- Duration and Air speed were still observed to have missing values.
- Since we are not yet sure how critical duration is to the landing distance, removing the entire set of data for a missing duration might lead to loss of significant information.
- For air speed, once the distributions were plotted, it became evident that the speed has a truncated distribution with a clear lower bound of 90, implying that the observations below that threshold were not being observed for a reason rather than randomly being missing/not being collected. Thus, the air speed values were concluded to be critical and were not to be removed.
- Other variable distributions were also observed and were largely normal
- Correlation between variables was also looked at and for low correlations, plots were drawn to spot any non-linear correlation
- Since passenger count and duration didn't have much correlation with distance, they were dropped from the data
- Prior to modeling, it was to be tested if the aircraft class has any impact on distance. A t test was run to test the difference in means of the distance and it was proved that it has a significant impact. As a result of which, the data was imputed with a dummy variable for the airline class.

Modeling

- In the first run, with pitch included, since the coefficient could not be concluded as non-zero, regression was rerun with that variable excluded.

- A choice of 2 models is being presented herewith to the end user. This is to not discard the relatively few air speed values.
- One model will be used when speed air is captured, other when speed air is missing. If the end user wants a more simplistic approach, the model with speed ground could be used. This has a slightly lower R squared but not significantly different. The 2 or 1 model approach appears to be a better option than imputing the air speed values.
- Since air speed had higher correlation to the distance, when it was present and was also resulting in much higher R squared, the speed ground was ignored since it would cause multicollinearity.

SAS Code:

```
/*1-Importing Input files FAA1 and FAA2*/
FILENAME REFFILE '/home/satyasmc0/Stat computing class/FAA1.xls';
PROC IMPORT DATAFILE=REFFILE
        DBMS=XLS
        OUT=WORK.faa1;
        GETNAMES=YES;
        sheet=FAA1;
run;

FILENAME REFFILE '/home/satyasmc0/Stat computing class/FAA2.xls';
PROC IMPORT DATAFILE=REFFILE
        DBMS=XLS
        OUT=WORK.faa2;
        GETNAMES=YES;
        sheet=faa2;

proc means data=faa1 n mean std range min max nmiss;
proc means data=faa2  n mean std range min max nmiss;
run;

/*2-from the means procedure, it is evident that the faa1 and faa2 belong to the same population
thus data can be combined*/
data faa3;
set faa1 faa2;
proc means data=faa3 n mean std range min max nmiss;
run;

/*3- Checking for duplicates*/
proc sort data= faa3 nodupkey
out = faa4;
by pitch;
proc print data=faa4;
run;
```

```
/*4- Removing missing values*/
data faa4;
set faa4;
if missing(aircraft) then delete;
run;
proc sort data=faa4;
by aircraft;
proc means data=faa4 n mean std range min max nmiss;
title 'All data summary- unique';
run;
/*No missing values found in summary- duration is missing for 50 values*/

/*5- Checking and removing abnormal values*/
data faa_normal;
set faa4;

if duration=. then miss='yes';
if speed_ground=. then miss='yes';
if speed_air=. then miss='yes';
if height=. then miss='yes';
if distance=. then miss='yes';
if pitch=. then miss='yes';

if duration<=40 and duration <> . then abnormal='yes';
if speed_ground<30 or speed_ground>140 and speed_ground <> . then abnormal='yes';
if (speed_air<30 or speed_air>140) and speed_air <> . then abnormal='yes';
if height<6 and height <> . then abnormal='yes';
if distance>6000 and distance <> . then abnormal='yes';
run;
proc sort data=faa_normal;
by abnormal miss;
proc print data=faa_normal;
proc means data = faa_normal n nmiss min max  ;
run;

/*Since abnormal values are a very small percentage of the entire data, deleting them*/
data faa_normal;
set faa_normal;
if abnormal='yes' then delete;
drop abnormal;
drop miss;
proc means data = faa_normal n nmiss min max  ;
run;
/*We end up with 831 values with their summary*/

/* Comparing distributions indicates the speed_air is a truncated dist so better to seggregate*/
proc chart data= faa_normal;
vbar speed_air;
run;
proc chart data= faa_normal;
vbar speed_ground;
run;
```

```
data faa_normal;
set faa_normal;
if speed_air= . then Group = 0; else Group = 1;
proc print data=faa_normal;
proc means data = faa_normal n nmiss min max;
run;

/*Exploring Variable distributions*/
proc univariate data=faa_normal;
class group;
var speed_ground;
histogram speed_ground;

proc univariate data = faa_normal;
class group;
var height;
histogram height;
proc univariate data = faa_normal;
class group;
var pitch;
histogram pitch;
proc univariate data = faa_normal;
class group;
var no_pasg;
histogram no_pasg;
/*Height, pitch and passenger count variables are nearly normal*/

/*Exploring Variable Correlations*/
proc corr data=faa_normal;
var duration speed_air speed_ground no_pasg pitch height distance;
run;

/*The correlation matrix shows us that there is no impact of passenger count and duration on distance
Also as expected speed air and speed ground are heavily correlated*/

/*We can plot to see if there is any non-linear correlation*/
proc gplot ; plot distance*height;
proc gplot ; plot distance*pitch;
proc gplot ; plot distance*no_pasg;

/*Drop duration and passenger count*/
data faa_trim;
set faa_normal;
drop duration no_pasg;
run;
proc means data=faa_trim n nmiss min max;
run;

/*Impact of aircraft class*/
proc ttest data=faa_trim;
class aircraft;
var distance;
title 'Mean distance across Airbus and Boeing';
```

```
run;
/*pvalue<alpha so refer Satterthwaite section implying unequal variances*/
/*p value of ttest implies that the mean equality can be rejected
Created a dummy variable that can be used in regression*/

data faa_final;
set faa_trim;
if aircraft= 'boeing' then planetype = 0; else planetype = 1;
run;

/*Regressing the landing distance against variables
- building 2 models in order to not delete out the air speed*/
proc reg data = faa_final;
model distance = planetype speed_air pitch height;
title 'Regression when air speed is available';

proc reg data = faa_final;
model distance = planetype speed_ground pitch height;
title 'Regression when air speed is unavailable;
```

SAS Outputs:

1. Initial summary of data when it was loaded

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Range	Minimum	Maximum	N Miss
duration	duration	800	154.0065385	49.2592338	290.8575036	14.7642071	305.6217107	0
no_pasg	no_pasg	800	60.1325000	7.5271686	58.0000000	29.0000000	87.0000000	0
speed_ground	speed_ground	800	79.5414195	19.2348870	113.4829200	27.7357153	141.2186354	0
speed_air	speed_air	200	103.8294713	10.4118729	51.7220771	90.0028586	141.7249357	600
height	height	800	30.1217717	10.2761691	63.4922163	-3.5462524	59.9459639	0
pitch	pitch	800	4.0183751	0.5248160	3.6423041	2.2844801	5.9267842	0
distance	distance	800	1544.52	938.2330999	6498.97	34.0807833	6533.05	0

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Range	Minimum	Maximum	N Miss
no_pasg	no_pasg	150	60.3400000	7.3107717	34.0000000	44.0000000	78.0000000	50
speed_ground	speed_ground	150	77.9173910	19.8788997	111.9909790	29.2276564	141.2186354	50
speed_air	speed_air	39	103.2224489	11.6781942	51.6139224	90.1110133	141.7249357	161
height	height	150	30.2326030	10.8272955	61.6297972	-3.5462524	58.0835448	50
pitch	pitch	150	4.0238987	0.5342237	2.8874935	2.6689057	5.5563992	50
distance	distance	150	1571.77	1005.55	6107.19	425.8585610	6533.05	50

2. Data summary once duplicates were removed

All data summary- unique

The MEANS Procedure

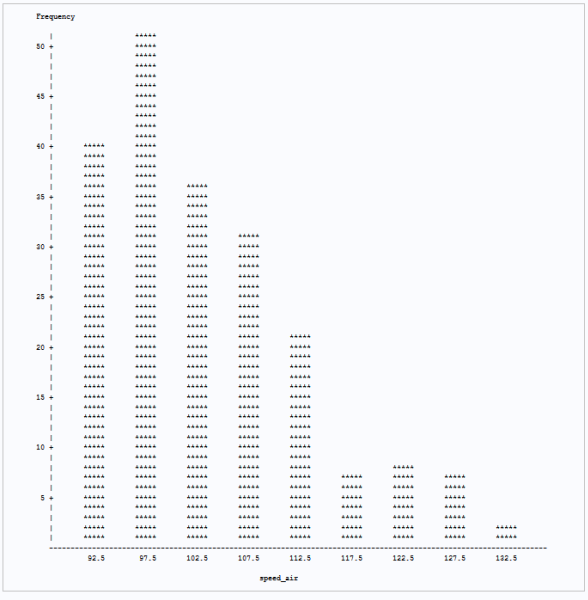
Variable	Label	N	Mean	Std Dev	Range	Minimum	Maximum	N Miss
duration	duration	800	154.0065385	49.2592338	290.8575036	14.7642071	305.6217107	50
no_pasg	no_pasg	850	60.1035294	7.4931370	58.0000000	29.0000000	87.0000000	0
speed_ground	speed_ground	850	79.4523229	19.0594903	113.4829200	27.7357153	141.2186354	0
speed_air	speed_air	208	103.7977237	10.2590370	51.7220771	90.0028586	141.7249357	642
height	height	850	30.1442223	10.2877268	63.4922163	-3.5462524	59.9459639	0
pitch	pitch	850	4.0093577	0.5288298	3.6423041	2.2844801	5.9267842	0
distance	distance	850	1526.02	928.5600816	6498.97	34.0807833	6533.05	0

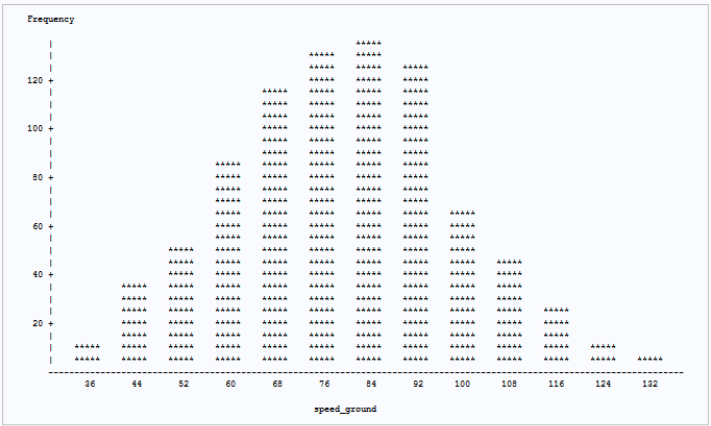
3. Summary when abnormal values were removed- 831 values remain

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum
duration	duration	781	50	41.9493694	305.6217107
no_pasg	no_pasg	831	0	29.0000000	87.0000000
speed_ground	speed_ground	831	0	33.5741041	132.7846766
speed_air	speed_air	203	628	90.0028586	132.9114649
height	height	831	0	6.2275178	59.9459639
pitch	pitch	831	0	2.2844801	5.9267842
distance	distance	831	0	41.7223127	5381.96

4. Since air speed values had most blanks, examining the distributions of air speed and air ground



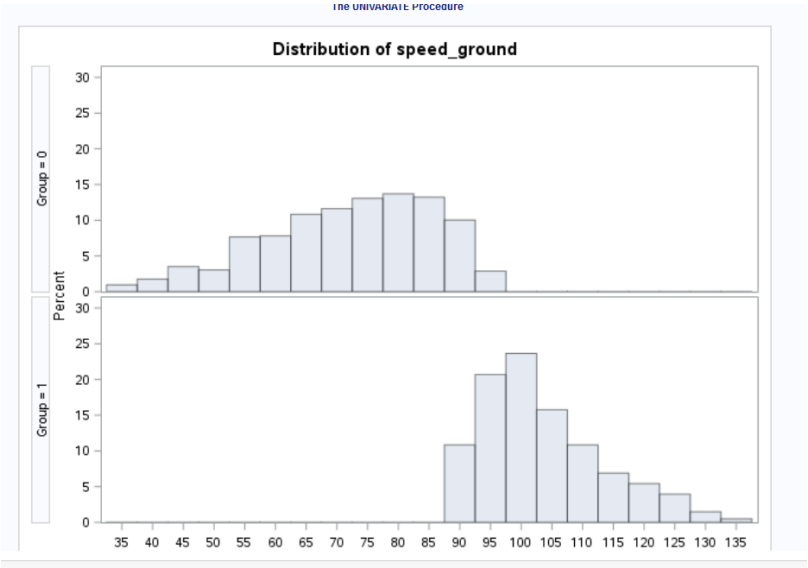


5. Since air speed values were to be retained for the model, created two groups

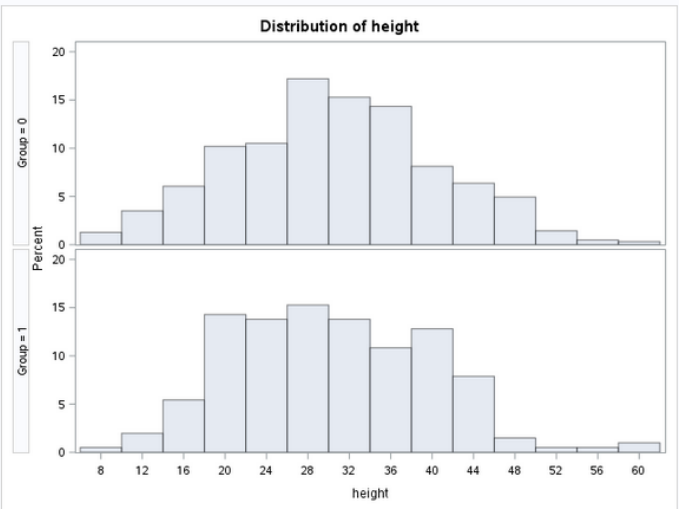
The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum
duration	duration	781	50	41.9493694	305.6217107
no_pasg	no_pasg	831	0	29.0000000	87.0000000
speed_ground	speed_ground	831	0	33.5741041	132.7846766
speed_air	speed_air	203	628	90.0028586	132.9114649
height	height	831	0	6.2275178	59.9459639
pitch	pitch	831	0	2.2844801	5.9267842
distance	distance	831	0	41.7223127	5381.96
Group		831	0	0	1.0000000

6. Creating groups made the difference in distributions very clear for speed ground, while the rest were largely normal.



The UNIVARIATE Procedure



7. Understanding linear correlations

The CORR Procedure

7 Variables:	duration speed_air speed_ground no_pasg pitch height distance

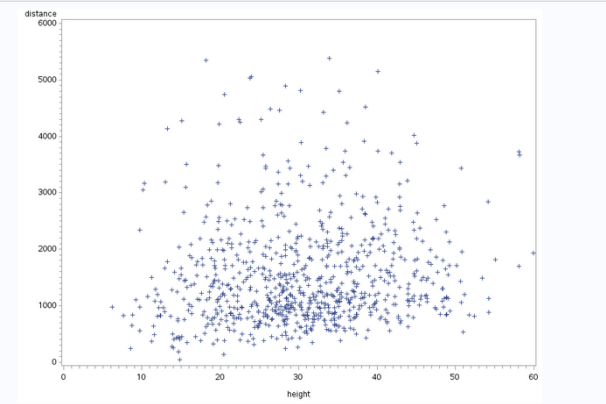
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
duration	781	154.77572	48.34992	120880	41.94937	305.62171	duration
speed_air	203	103.48504	9.73628	21007	90.00286	132.91146	speed_air
speed_ground	831	79.54270	18.73568	66100	33.57410	132.78468	speed_ground
no_pasg	831	60.05535	7.49132	49906	29.00000	87.00000	no_pasg
pitch	831	4.00516	0.52657	3328	2.28448	5.92678	pitch
height	831	30.45787	9.78481	25310	6.22752	59.94596	height
distance	831	1522	896.33815	1265183	41.72231	5382	distance

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

	duration	speed_air	speed_ground	no_pasg	pitch	height	distance
duration duration	1.00000 781	0.04454 0.5364 195	-0.04897 0.1716 781	-0.03639 0.3098 781	-0.04675 0.1918 781	0.01112 0.7564 781	-0.05138 0.1514 781
speed_air speed_air	0.04454 0.5364 195	1.00000 203	0.98794 <.0001 203	-0.00616 0.9305 203	-0.03927 0.5780 203	-0.07933 0.2606 203	0.94210 <.0001 203
speed_ground speed_ground	-0.04897 0.1716 781	0.98794 <.0001 203	1.00000 831	-0.00013 0.9969 831	-0.03912 0.2599 831	-0.05761 0.0970 831	0.86624 <.0001 831
no_pasg no_pasg	-0.03639 0.3098 781	-0.00616 0.9305 203	-0.00013 0.9969 831	1.00000 831	-0.01793 0.6057 831	0.04699 0.1760 831	-0.01776 0.6093 831
pitch pitch	-0.04675 0.1918 781	-0.03927 0.5780 203	-0.03912 0.2599 831	-0.01793 0.6057 831	1.00000 831	0.02298 0.5082 831	0.08703 0.0121 831
height height	0.01112 0.7564 781	-0.07933 0.2606 203	-0.05761 0.0970 831	0.04699 0.1760 831	0.02298 0.5082 831	1.00000 831	0.09941 0.0041 831
distance distance	-0.05138 0.1514 781	0.94210 <.0001 203	0.86624 <.0001 831	-0.01776 0.6093 831	0.08703 0.0121 831	0.09941 0.0041 831	1.00000 831

8. Checking for non-linear correlation



9. Dropping duration and no of passengers due to poor correlation

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum
speed_ground	speed_ground	831	0	33.5741041	132.7846766
speed_air	speed_air	203	628	90.0028586	132.9114649
height	height	831	0	6.2275178	59.9459639
pitch	pitch	831	0	2.2844801	5.9267842
distance	distance	831	0	41.7223127	5381.96
Group		831	0	0	1.0000000

10. Examining impact of aircraft – Boeing and airbus: concluded there is a difference

Mean distance across Airbus and Boeing

The TTEST Procedure
Variable: distance (distance)

aircraft	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus	444	1323.3	791.9	37.5833	41.7223	4896.3
boeing	387	1751.0	953.9	48.4869	573.6	5382.0
Diff (1-2)		-427.7	871.1	60.5772		

aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		1323.3	1249.5	1397.2	791.9	743.0	847.8
boeing		1751.0	1655.7	1846.3	953.9	891.1	1026.2
Diff (1-2)	Pooled	-427.7	-546.6	-308.8	871.1	831.1	915.1
Diff (1-2)	Satterthwaite	-427.7	-548.1	-307.2			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	829	-7.06	<.0001
Satterthwaite	Unequal	752.49	-6.97	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	386	443	1.45	0.0002

11. Created dummy variable for aircraft and ran correlation check

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
speed_ground	speed_ground	831	79.5426997	18.7356754	33.5741041	132.7846766
speed_air	speed_air	203	103.4850352	9.7362774	90.0028586	132.9114649
height	height	831	30.4578695	9.7848114	6.2275178	59.9459639
pitch	pitch	831	4.0051609	0.5265690	2.2844801	5.9267842
distance	distance	831	1522.48	896.3381524	41.7223127	5381.96
Group		831	0.2442840	0.4299206	0	1.0000000
planetype		831	0.5342960	0.4991228	0	1.0000000

12. Regression for the group where air speed is available

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	203
Number of Observations with Missing Values	628

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	132881890	33220473	1834.22	<.0001
Error	198	3586078	18112		
Corrected Total	202	136467968			

Root MSE	134.57899	R-Square	0.9737
Dependent Mean	2774.67289	Adj R-Sq	0.9732
Coeff Var	4.85027		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5955.30400	133.31678	-44.67	<.0001
planetype		1	-428.11078	20.39889	-20.99	<.0001
speed_air	speed_air	1	82.14684	0.97860	83.94	<.0001
pitch	pitch	1	-1.75734	17.93681	-0.10	0.9221
height	height	1	13.69957	1.00991	13.57	<.0001

13.Regressing with ground speed

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	566395312	141598828	1164.42	<.0001
Error	826	100445017	121604		
Corrected Total	830	666840329			

Root MSE	348.71785	R-Square	0.8494
Dependent Mean	1522.48287	Adj R-Sq	0.8486
Coeff Var	22.90455		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2183.05415	123.64155	-17.66	<.0001
planetype		1	-481.26818	25.95117	-18.55	<.0001
speed_ground	speed_ground	1	42.42833	0.64788	65.49	<.0001
pitch	pitch	1	39.60761	24.59908	1.61	0.1078
height	height	1	14.09086	1.23977	11.37	<.0001

Since pitch is shown to have 0 coefficient in both cases(high p value), rerunning the regression without pitch

14. Revised outputs when pitch is removed

Regression when air speed is available

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	203
Number of Observations with Missing Values	628

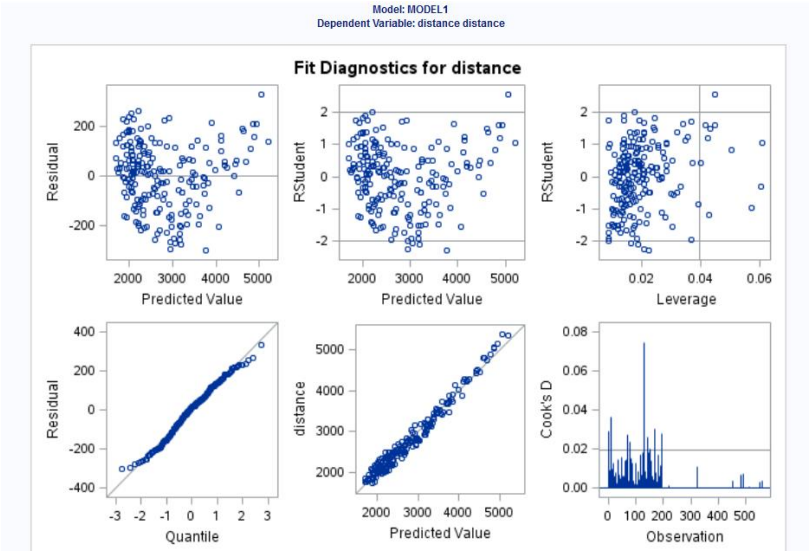
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	132881716	44293905	2457.85	<.0001
Error	199	3586252	18021		
Corrected Total	202	136467968			

Root MSE	134.24368	R-Square	0.9737
Dependent Mean	2774.67289	Adj R-Sq	0.9733
Coeff Var	4.83818		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5962.93395	107.93619	-55.24	<.0001
planetype		1	-427.44156	19.17339	-22.29	<.0001
speed_air	speed_air	1	82.14852	0.97601	84.17	<.0001
height	height	1	13.70161	1.00718	13.60	<.0001



15. Final output of regression with ground speed

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	566080053	188693351	1548.72	<.0001
Error	827	100760276	121838		
Corrected Total	830	666840329			

Root MSE	349.05344	R-Square	0.8489
Dependent Mean	1522.48287	Adj R-Sq	0.8484
Coeff Var	22.92659		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2016.19809	67.50541	-29.87	<.0001
planetype		1	-496.04524	24.29753	-20.42	<.0001
speed_ground	speed_ground	1	42.40242	0.64830	65.41	<.0001
height	height	1	14.14783	1.24046	11.41	<.0001