BANA 7042 PROJECT

Name: Syed Imad Husain **UCID: M12958531**

Contents

1. Introduction

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).

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.

Part 1. Practice of modeling the landing distance using linear regression.

Please write R programs to complete the following steps. In each step, provide

- The R code (how do you realize it?)
- The R output (Copy and paste only those relevant)
- Your observations (What do you observe from the output?)
- Your conclusion/decision

2. Initial exploration of the data

Step 1. Read the two files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' into your R system. Please search "Read Excel files from R" in Google in case you do not know how to do that.

- **Code**

```
library(readxl)
FAA1 <- read_excel("FAA1.xls")
FAA2 <- read_excel("FAA2.xls")
```

- **Relevant Output**

NA

- **Observations**

NA

- **Conclusion**

NA

Step 2. Check the structure of each data set using the "str" function. For each data set, what is the sample size and how many variables? Is there any difference between the two data sets?

- **Code**

```
str(FAA1)
str(FAA2)
variable.names(FAA1)
variable.names(FAA2)
setdiff(variable.names(FAA1),variable.names(FAA2))
```

- **Relevant Output**

```
FAA1 800 obs. of  8 variables:
 $ aircraft    : chr  "boeing" "boeing" "boeing" "boeing" ...
 $ duration    : num  98.5 125.7 112 196.8 90.1 ...
 $ no_pasg     : num  53 69 61 56 70 55 54 57 61 56 ...
 $ speed_ground: num  107.9 101.7 71.1 85.8 59.9 ...
 $ speed_air   : num  109 103 NA NA NA ...
 $ height      : num  27.4 27.8 18.6 30.7 32.4 ...
 $ pitch       : num  4.04 4.12 4.43 3.88 4.03 ...
 $ distance    : num  3370 2988 1145 1664 1050 ...
FAA2 150 obs. of  7 variables:
 $ aircraft    : chr  "boeing" "boeing" "boeing" "boeing" ...
 $ no_pasg     : num  53 69 61 56 70 55 54 57 61 56 ...
 $ speed_ground: num  107.9 101.7 71.1 85.8 59.9 ...
 $ speed_air   : num  109 103 NA NA NA ...
 $ height      : num  27.4 27.8 18.6 30.7 32.4 ...
 $ pitch       : num  4.04 4.12 4.43 3.88 4.03 ...
 $ distance    : num  3370 2988 1145 1664 1050 ...
```

Variables in FAA1 that are not present in FAA2 "duration"

- **Observations**

FAA1 has 800 observations and 8 variables. FAA2 has 150 observations but only 7 variables.

- **Conclusion**

The variable Duration is not present in FAA2

Step 3. Merge the two data sets. Are there any duplications? Search "check duplicates in r" if you do not know how to check duplications. If the answer is "Yes", what action you would take?

- **Code**

```
#remove duplicates if any within data set
FAA1 <- unique(FAA1)
FAA2 <- unique(FAA2)

# check for duplicates between the data sets
merged <- rbind(FAA1[,-2],FAA2)
sum(duplicated(merged))

#finally create unique dataset
merged <- unique(merged)
FAA <- merge(merged,FAA1,by=names(merged), all.x=TRUE)
```

- **Relevant Output**

NA

- **Observations**

We observe that there are 100 duplicate values between FAA1 and FAA2

- **Conclusion**

We merge the dataset using MERGE command which does a left outer join between the two datasets based on the key columns

Step 4. Check the structure of the combined data set. What is the sample size and how many variables? Provide summary statistics for each variable.

- **Code**

```
str(FAA)
summary(FAA)
mat <- var(FAA[,-1],na.rm = T)
for(i in 1:7)   {
  print(paste(names(mat[,i])[i],mat[i,i]))
   }
table(FAA$aircraft)
```

- **Relevant Output**

850 obs. of 8 variables Summary

- **Observations**

Statistic	no_pasg	speed_ground	speed_air	height	pitch	distance	duration
Minimum	29	27.74	90	-3.546	2.284	34.08	14.76
1st Quartile	55	65.9	96.25	23.314	3.642	883.79	119.49
Median	60	79.64	101.15	30.093	4.008	1258.09	153.95
Mean	60.1	79.45	103.8	30.144	4.009	1526.02	154.01
3rd Quartile	65	92.06	109.4	36.993	4.377	1936.95	188.91
Maximum	87	141.22	141.72	59.946	5.927	6533.05	305.62
Missing Values	0	0	642	0	0	0	50
Standard Deviation	7.006	10.572	10.412	9.550	0.563	899.471	50.209

Table 1 - Summary Statistics for Numerical Variables

Sr. No.	Values	Count
1	Airbus	450
2	boeing	400

Table 2 - Summary Statistics for Categorical Variables

- **Conclusion**

There seems to be some outliers and abnormal values in the data and hence we need to perform outlier treatment

Step 5. By now, if you are asked to prepare ONE presentation slide to summarize your findings, what observations will you bring to the attention of FAA agents?

- There were few duplicate values and there are certain abnormal values which do not meet the criteria per the data dictionary
- Is it possible to obtain data for the variables Speed Air and Duration because there are a lot of missing values?
- When was this data collected? Is it too old? What is the span over which the data?
- Where are airports located? Is it just one Airport? Where will the outcome of this project be applied?
- Is it possible to gather data on parameters like time, day, date, weather, etc. during data collection? Is there anymore data related to flights that is available like Pilot, Copilot profiles, etc.?

Please list no more than five using “bullet statements”, from the most important to the least important.

3. Data Cleaning and further exploration

Step 6. Are there abnormal values in the data set? Please refer to the variable dictionary for criteria defining "normal/abnormal" values. Remove the rows that contain any "abnormal values" and report how many rows you have removed.

- **Code**

```
faa_clean <- FAA %>% select(names(FAA)) %>%
  filter(replace(duration,is.na(duration),60) > 40 &
          (speed_ground > 30 &
          speed_ground < 140) &
          (replace(speed_air,is.na(speed_air),60) > 30 &
          replace(speed_air,is.na(speed_air),60) < 140) &
          height >= 6 &
          distance < 6000
         )
```

- **Relevant Output**

NA

- **Observations**

Following cleaning rules were applied –
duration > 40
30 < speed ground < 140
30 < speed air < 140
height > 6
distance < 6000

- **Conclusion**

Finally, after removing abnormal values, we are left with 831 rows

Step 7. Repeat Step 4.

- **Code**

```
str(faa_clean)
summ <- summary(faa_clean)
mat <- sqrt(var(faa_clean[,-1],na.rm = T))
for(i in 1:7)   {
  print(paste(names(mat[,i])[i],mat[i,i]))
}
table(faa_clean$aircraft)
```

- **Relevant Output**

NA

- **Observations**

Statistics	no_pasg	speed_ground	speed_air	height	pitch	distance	duration
Minimum	29	33.57	90	6.228	2.284	41.72	41.95
1st Quartile	55	66.2	96.23	23.53	3.64	893.28	119.63
Median	60	79.79	101.12	30.167	4.001	1262.15	154.28
Mean	60.06	79.54	103.48	30.458	4.005	1522.48	154.78
3rd Quartile	65	91.91	109.36	37.004	4.37	1936.63	189.66
Maximum	87	132.78	132.91	59.946	5.927	5381.96	305.62
Missing Values			628				50
Standard Deviation	7.03675	10.05138332	9.880376	9.3906	0.561	830.091	48.1531

Table 3 Summary of Numerical Variables

Sr. No.	Values	Count
1	Airbus	444
2	boeing	387

Table 4 Summary of Categorical Variables

- **Conclusion**

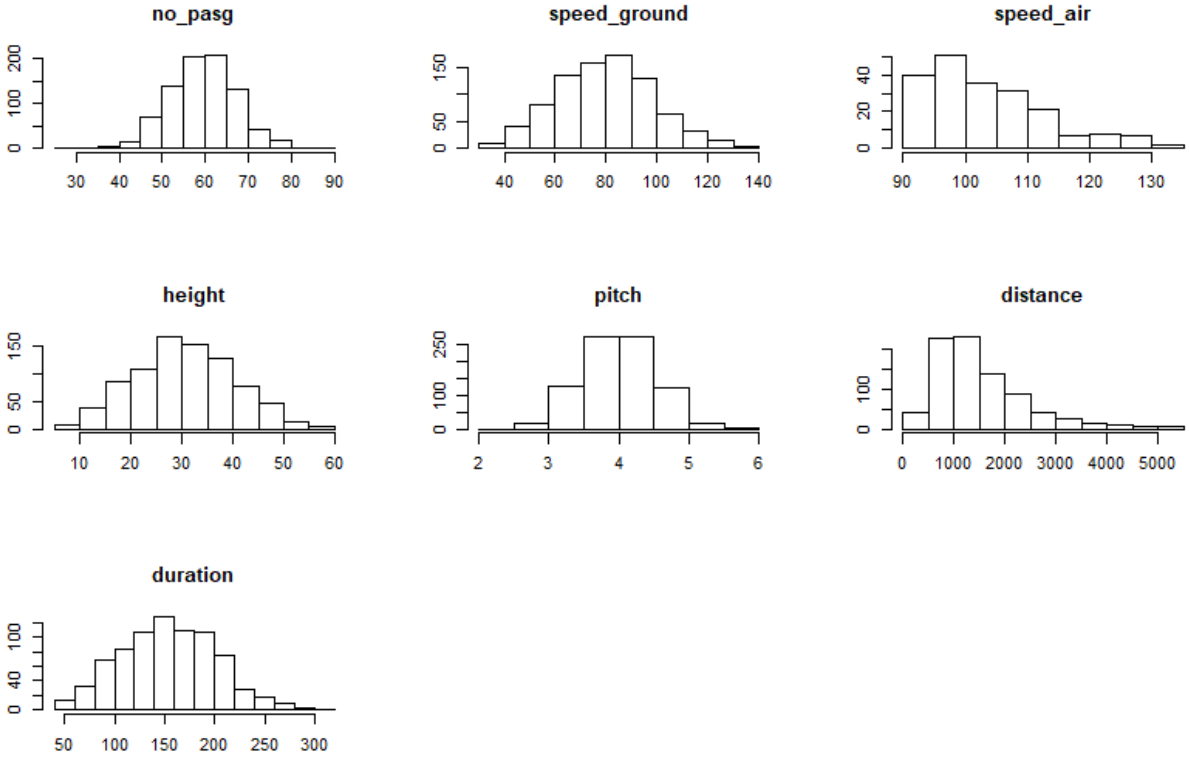
The Summary statistics have changed as compared to earlier and do not reveal any apparent anomalies with the data

Step 8. Since you have a small set of variables, you may want to show histograms for all of them.

- **Code**

```
par(mfrow = c(3,3))
for(i in 2:8) {
hist(faa_clean[,i],main = names(faa_clean)[i],xlab = '',ylab='') }
```

- **Relevant Output**



- **Observations**

Most of the variables appear to have a normal distribution. However, Distance and Speed Air appear to be right skewed because there are many outliers

- **Conclusion**

We conclude that Distance and Speed Air have outliers and all variables can be approximated to a normal distribution

Step 9. Prepare another presentation slide to summarize your findings drawn from the cleaned data set, using no more than five "bullet statements".

- The cleaned data set after removing abnormal values has 831 rows and 8 variables
- Observations where Duration or Speed Air were missing were not removed
- Distance being the response variable has many outliers on the higher end as observed from the Histogram
- Speed Air has a truncated left tail as all available values are >90 mph. However, it has a heavy right tail indicating outliers
- All variables can e approximated to a normal distribution

4. Initial analysis for identifying important factors that impact the response variable "landing distance"

Step 10. Compute the pairwise correlation between the landing distance and each factor X. Provide a table that ranks the factors based on the size (absolute value) of the correlation. This table contains three columns: the names of variables, the size of the correlation, the direction of the correlation (positive or negative). We call it Table 1, which will be used for comparison with our analysis later.

- **Code**

```
cor(faa_clean[,-1],use = "complete.obs")[,6]
```

- **Relevant Output**

Variable	Correlation	Sign
no_pasg	0.033	Negative
speed_ground	0.929	Positive
speed_air	0.943	Positive
height	0.058	Positive
pitch	0.034	Positive
duration	0.052	Positive

Table 5- Correlation Table

- **Observations**

Only Number of Passengers has Negative correlation. Only Speed Ground and Speed Air have strong correlation

- **Conclusion**

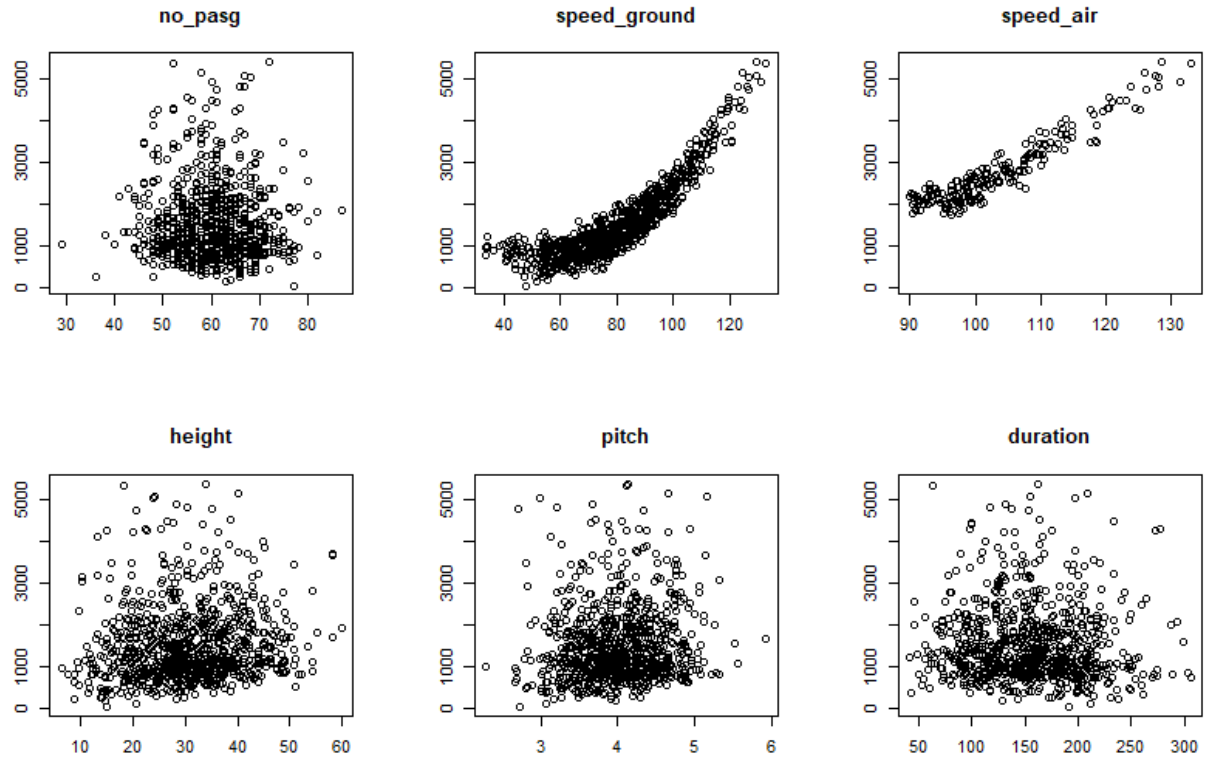
Not all variables seem to have an effective correlation. Only Speed Ground and Speed Air seem to be significant at this point in time

Step 11. Show X-Y scatter plots. Do you think the correlation strength observed in these plots is consistent with the values computed in Step 10?

- **Code**

```
par(mfrow = c(2,3))
for(i in c(2,3,4,5,6,8)) {
plot(faa_clean[,i],faa_clean$distance ,main = names(faa_clean)[i],
     xlab = '',ylab='') }
```

- **Relevant Output**



- **Observations**

Observable correlation of Distance can be seen with Speed Ground and Speed Air

- **Conclusion**

The results of Scatter plots are consistent with our findings in Step 10

Step 12. Have you included the airplane make as a possible factor in Steps 10-11? You can code this character variable as 0/1.

- **Code**

```
faa_clean$airbus <- 1
faa_clean$airbus[faa_clean$aircraft == "boeing"] <- 0
faa_clean$airbus <- as.factor(faa_clean$airbus)
```

- **Relevant Output**

NA

- **Observations**

The added variable has same summary statistics as the variable aircraft which means our encoding is correct

- **Conclusion**

We have added a dummy encoding variable called 'airbus' which has a value of 1 when the aircraft is 'Airbus' else has a value of 0 indicating that the aircraft is 'Boeing'. Also we have encoded it as a Factor variable

5. Regression using a single factor each time

Step 13. Regress Y (landing distance) on each of the X variables. Provide a table that ranks the factors based on its significance. The smaller the p-value, the more significant the factor. This table contains three columns: the names of variables, the size of the p-value, the direction of the regression coefficient (positive or negative). We call it Table 2.

- **Code**

```
coeff <- rep (0,7)
p_val <- rep(0,7)
var_name <- rep('',7)
j <- 1
for(i in c(2,3,4,5,6,8,9)) {
fit <- lm(faa_clean[,7] ~ faa_clean[,i])
var_name[j] <- names(faa_clean)[i]
p_val[j] <- summary(fit)$coefficients[2,4]
coeff[j] <- summary(fit)$coefficients[2,1]
j <- j+1
}
tt <- cbind(var_name,coeff,p_val)
```

- **Relevant Output**

Factor	Sign	P Value
speed_ground	Positive	4.77E-252
speed_air	Positive	2.50E-97
airbus	Negative	3.53E-12
height	Positive	0.00412386
pitch	Positive	0.012081242
duration	Negative	0.151400209
no_pasg	Negative	0.609252002

Table 6 Regression Table

- **Observations**

Not all variables are significant as per the observations. Also, Airbus i.e. our dummy encoded variable has a negative coefficient along with duration and no of passenger

- **Conclusion**

Speed Ground, Speed Air, Airbus and Height are statistically significant in terms of the regression coefficients

Step 14. Standardize each X variable. In other words, create a new variable

X'= {X-mean(X)}/sd(X).

The mean of X' is 0 and its standard deviation is 1.
Regress Y (landing distance) on each of the X' variables. Provide a table that ranks the factors based on the size of the regression coefficient. The larger the size, the more important the factor. This table contains three columns: the names of variables, the size of the regression coefficient, the direction of the regression coefficient (positive or negative). We call it Table 3.

- **Code**

```
faa_std <- bind_cols(list(faa_clean[,1],
                  as.numeric(round((faa_clean[,2]-mean(faa_clean[,2],na.omit =
T))/sd(faa_clean[,2],na.rm=T),3)),
                  as.numeric(round((faa_clean[,3]-mean(faa_clean[,3],na.omit =
T))/sd(faa_clean[,3],na.rm=T),3)),
                  as.numeric(round((faa_clean[,4]-mean(faa_clean[,4],na.rm =
T))/sd(faa_clean[,4],na.rm=T),3)),
                  as.numeric(round((faa_clean[,5]-mean(faa_clean[,5],na.omit =
T))/sd(faa_clean[,5],na.rm=T),3)),
                  as.numeric(round((faa_clean[,6]-mean(faa_clean[,6],na.omit =
T))/sd(faa_clean[,6],na.rm=T),3)),
                  as.numeric(faa_clean[,7]),
                  as.numeric(round((faa_clean[,8]-mean(faa_clean[,8],na.rm =
T))/sd(faa_clean[,8],na.rm=T),3)),
                  faa_clean[,9]))

names(faa_std) <- names(faa_clean)

table(faa_std[,9])
table(faa_clean[,9])
str(faa_std)
coeff <- rep (0,7)
p_val <- rep(0,7)
var_name <- rep('',7)
j <- 1
for(i in c(2,3,4,5,6,8,9)) {
  fit <- lm(faa_std[,7] ~ faa_std[,i])
  var_name[j] <- names(faa_std)[i]
  p_val[j] <- summary(fit)$coefficients[2,4]
  coeff[j] <- summary(fit)$coefficients[2,1]
  j <- j+1
}
tt1 <- cbind(var_name,coeff,p_val)
```

- **Relevant Output**

var_name	coeff	p_val
speed_ground	776.441	4.85E-252
speed_air	774.349	2.50E-97
airbus	427.666	3.53E-12
height	89.1048	0.004124358
pitch	78.025	0.012061172
duration	46.472	0.151462938
no_pasg	15.9138	0.609303679

Table 7Regression after Standardization

- **Observations**

The that the data has been standardized, not all variables are significant as per the observations.

- **Conclusion**

Speed Ground, Speed Air and Airbus remain the most significant variables in terms of Coefficients' size

Step 15. Compare Tables 1,2,3. Are the results consistent? At this point, you will meet with a FAA agent again. Please provide a single table than ranks all the factors based on their relative importance in determining the landing distance. We call it Table 0.

- **Observations**

Variable	Table 1: Correlation	Table 2: P Value Regression	Table 3: Standardized Coefficients
speed_ground	2	1	**1**
speed_air	1	2	**2**
airbus	3	3	**3**
height	4	4	**4**
pitch	6	5	**5**
duration	5	6	**6**
no_pasg	7	7	**7**

Table 8 Factor Ranking - "Table 0"

We observe that the values in Table 2 and Table 3 are ranked the same. However, the ranks allotted by Table 1 i.e. Correlation, are slightly different

- **Conclusion**

Variable	Final Rank
speed_ground	1
speed_air	2
airbus	3
height	4
pitch	5
duration	6
no_pasg	7

Table 9 Final Ranking for Factors

We can check with the Agent based on his domain knowledge if this ranking is coherent.

6. Check collinearity

Step 16. Compare the regression coefficients of the three models below:

Model 1: LD ~ Speed_ground

Model 2: LD ~ Speed_air

Model 3: LD ~ Speed_ground + Speed_air
Do you observe any significance change and sign change? Check the correlation between Speed_ground and Speed_air. You may want to keep one of them in the model selection. Which one would you pick? Why?

- **Code**

```
LD <- faa_clean$distance
Speed_ground <- faa_clean$speed_ground
Speed_air <- faa_clean$speed_air
Model1 <- lm(LD ~ Speed_ground)
Model2 <- lm(LD ~ Speed_air)
Model3 <- lm(LD ~ Speed_ground + Speed_air)

summary(Model1)$coefficients
summary(Model2)$coefficients
summary(Model3)$coefficients
scatterplot(faa_std$speed_ground,faa_std$speed_air)
cor(faa_std$speed_ground,faa_std$speed_air,use = "complete.obs")
```

- **Relevant Output**

Factor	Model 1		Model 2		Model 3	
	coef	p val	coef	p val	coef	p val
Speed Ground	41.442	4.7664E-252	NA	NA	-14.373	0.258477
Speed Air	NA	NA	79.5321	2.50046E-97	93.959	6.99E-12

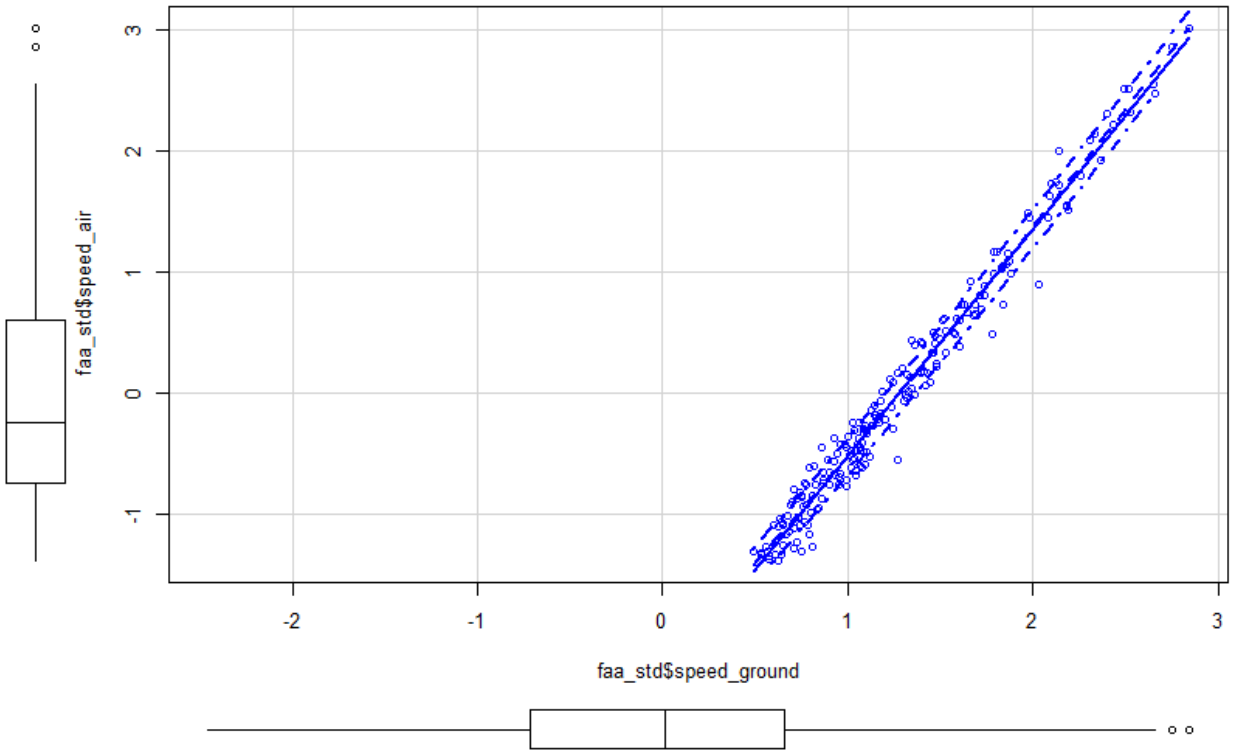


Table 10 Scatter plot between Speed Air and Speed Ground

```
Correlation Coef - 9879432
```

- **Observations**

The coefficient sign changes from model 1 to model 3. Also in Model 3, Speed Ground becomes statistically insignificant. Also, we visualize the correlation between the two variables in question and is is apparently very high. We performed a Correlation test between the two variables and the correlation is 0.98 which is very high

- **Conclusion**

We choose Speed Ground over Speed Air because we have to choose one. They do not differ much in statistical significance; however, we have more data for Speed Ground which since Speed Air has a lot of missing values

7. Variable selection based on our ranking in Table 0.

Step 17. Suppose in Table 0, the variable ranking is as follows: X1, X2, X3….. Please fit the following six models:

Model 1: LD ~ X1

Model 2: LD ~ X1 + X2

Model 3: LD ~ X1 + X2 + X3

…….

Calculate the R-squared for each model. Plot these R-squared values versus the number of variables p. What patterns do you observe?

- **Code**

```
Y <- faa_std$distance
X1 <- faa_std$speed_ground
X2 <- faa_std$airbus
X3 <- faa_std$height
X4 <- faa_std$pitch
X5 <- faa_std$duration
X6 <- faa_std$no_pasg

Model1 <- lm(Y ~ X1)
Model2 <- lm(Y ~ X1+X2)
Model3 <- lm(Y ~ X1+X2+X3)
Model4 <- lm(Y ~ X1+X2+X3+X4)
Model5 <- lm(Y ~ X1+X2+X3+X4+X5)
Model6 <- lm(Y ~ X1+X2+X3+X4+X5+X6)

r_sqaure <- c(
  summary(Model1)$r.squared,
  summary(Model2)$r.squared,
  summary(Model3)$r.squared,
  summary(Model4)$r.squared,
  summary(Model5)$r.squared,
  summary(Model6)$r.squared)

plot(r_sqaure,xlab='Number of Parameters',ylab='R Squared'
     ,main="Variation of R Squared with Number of Parameters")+
  lines.default(x = 1:6 , y = r_sqaure)
```

- **Relevant Output**

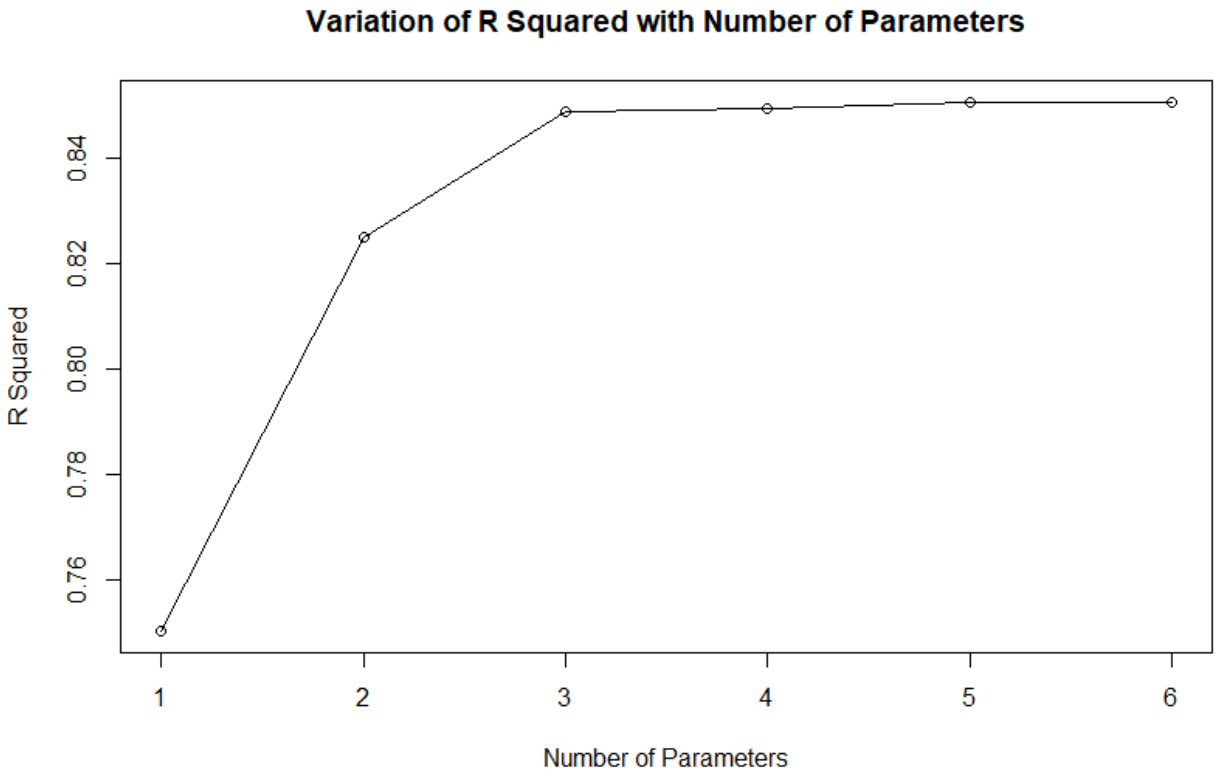


Table 11Variation of R squared against Number of Parameters

- **Observations**

We observe that there is a constant increase in the value of R squared as we keep on adding variables

- **Conclusion**

We conclude that R squared value keeps on increasing regardless of the statistical significance of the variable and hence it may not be the best of criterion to assess the optimality of a model

Step 18. Repeat Step 17 but use adjusted R-squared values

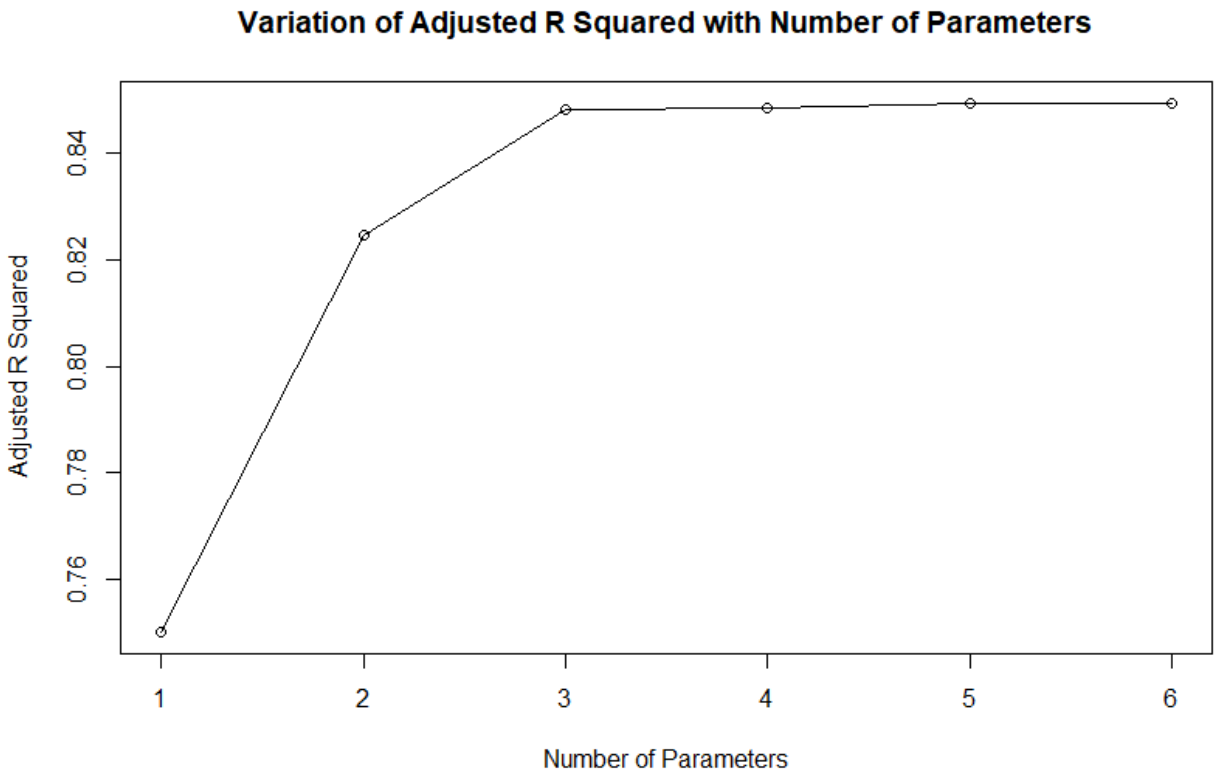
instead.

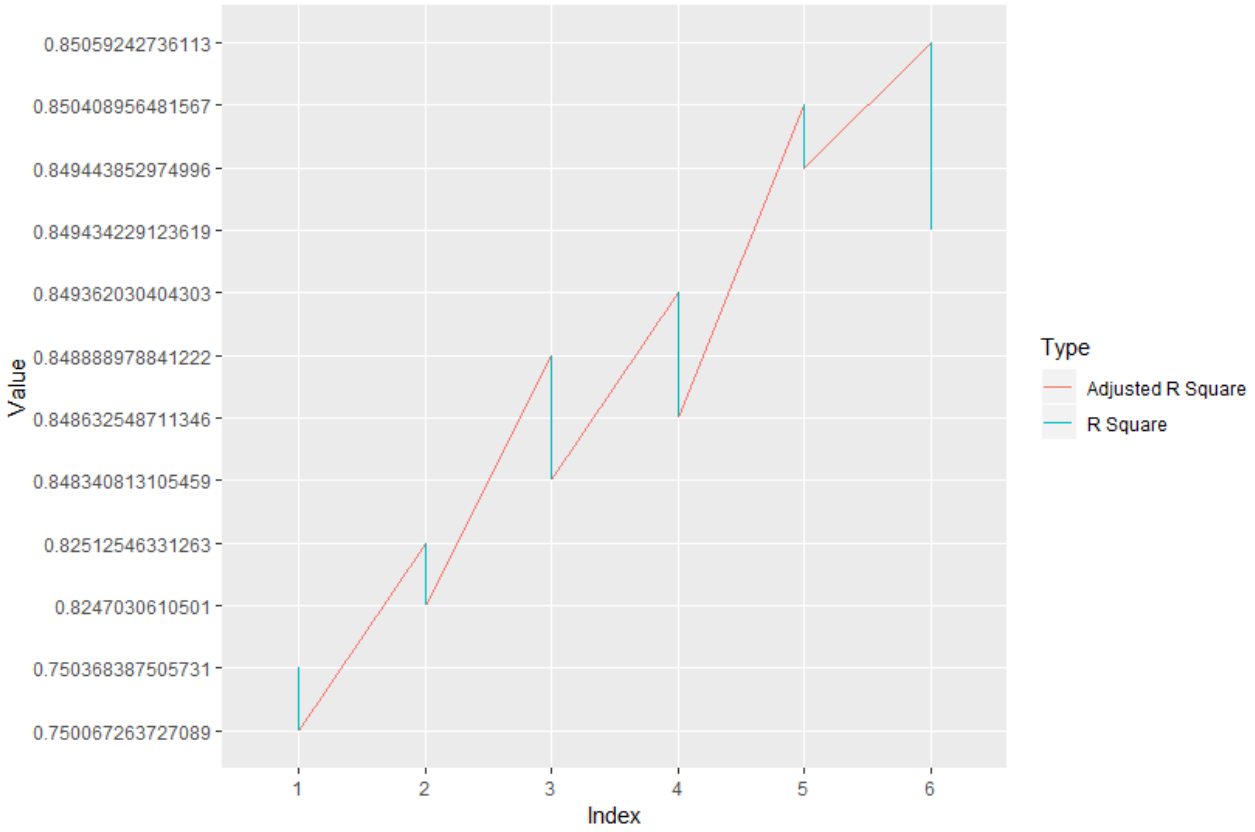
- **Code**

```
adj_r_sqr <- c(
  summary(Model1)$adj.r.squared,
  summary(Model2)$adj.r.squared,
  summary(Model3)$adj.r.squared,
  summary(Model4)$adj.r.squared,
  summary(Model5)$adj.r.squared,
  summary(Model6)$adj.r.squared)
```

```
plot(adj_r_sqr,xlab='Number of Parameters',ylab='Adjusted R Squared'
     ,main="Variation of Adjusted R Squared with Number of Parameters")+
  lines.default(x = 1:6 , y = adj_r_sqr)

R_sqr <- cbind(r_sqaure,"R Square",1:6)
Adj_R_sqr <- cbind(adj_r_sqr,"Adjusted R Square",1:6)

R_Cmp <- rbind(R_sqr,Adj_R_sqr)
R_Cmp <- as.data.frame(R_Cmp)

names(R_Cmp) <- c("Value","Type","Index")

install.packages("ggplot2")
library(ggplot2)


ggplot(data=R_Cmp,
       aes(x=Index, y=Value, colour=Type, group = 2)) +
  geom_line()
```

- **Relevant Output**





- **Observations**

We can observe that Adjusted R Square is increasing but not consistently

No. of Parameters	r_sqaure	adj_r_sqr
1	0.750368	0.750067
2	0.825126	0.824703
3	0.848889	0.848341
4	0.849362	0.848633
5	0.850409	0.849444
6	0.850592	0.849434

- **Conclusion**

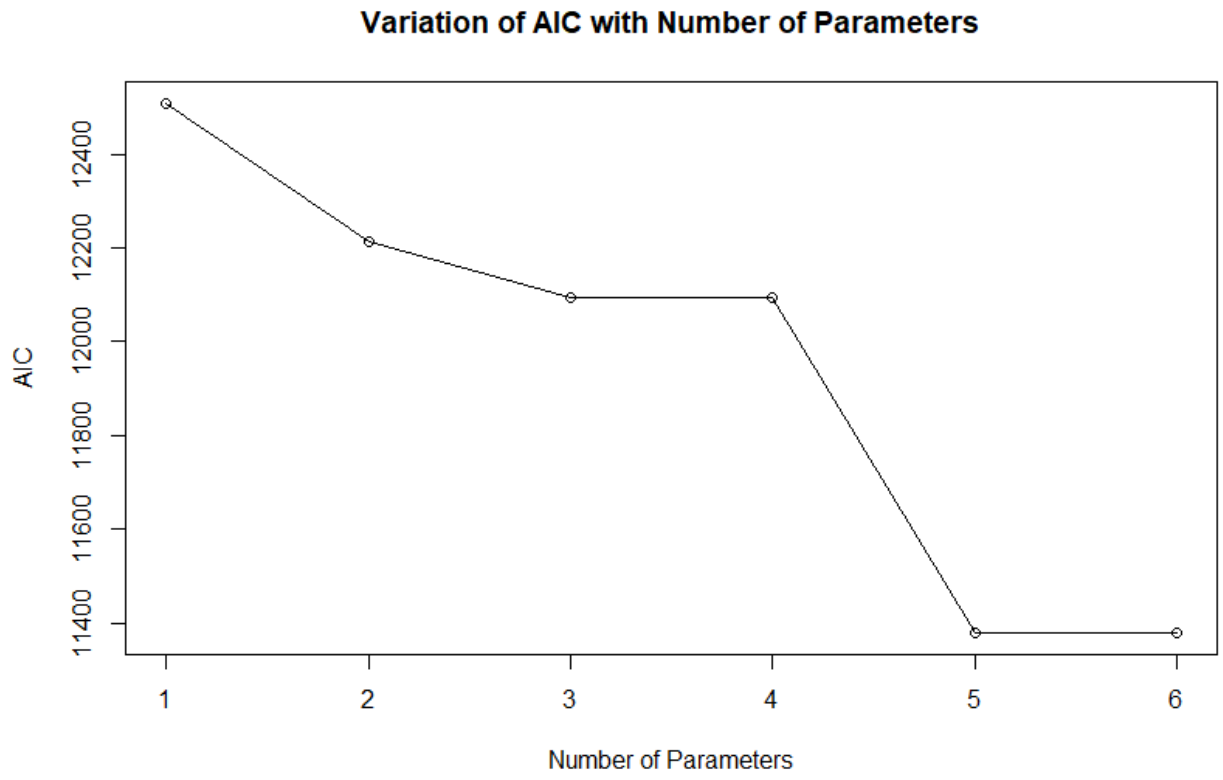
We conclude that Adjusted R square only increase when a variable has considerable significance. The green line in above's graph represents the distance between the two values at each index. We can see that Adjusted R Square is a better indicator of Model optimality as compared to R Squared

Step 19. Repeat Step 17 but use AIC values instead.

- **Code**

```
aic_cmp <- c(
  AIC(Model1),
  AIC(Model2),
  AIC(Model3),
  AIC(Model4),
  AIC(Model5),
  AIC(Model6))

plot(aic_cmp,xlab='Number of Parameters',ylab='AIC'
     ,main="Variation of AIC with Number of Parameters")+
  lines.default(x = 1:6 , y = aic_cmp)
```

- **Relevant Output**



- **Observations**

We notice that the value of AIC keeps on decreasing as variables are added, however the gradient is not always the same

No. of Parameters	AIC
1	12508.84
2	12215.08
3	12095.7
4	12095.1
5	11378.89
6	11379.93

- **Conclusion**

AIC is also an indicator to assess model optimality. As we can see, the AIC keeps on decreasing as we add more variables. As observed, model with 5 parameters has a lower value than the one with 6 parameters indicating that adding variables to the model does not reduce the AIC always but only when it is statistically significant.

Step 20. Compare the results in Steps 18-19, what variables would you select to build a predictive model for LD?

- **Observations**

No. of Parameters	R Squared	Adjusted R Squared	AIC
1	0.750368	0.7500673	12508.84
2	0.825126	0.8247031	12215.08
3	0.848889	0.8483408	12095.7
4	0.849362	0.8486325	12095.1
5	0.850409	0.8494439	11378.89
6	0.850592	0.8494342	11379.93

- **Conclusion**

The best Model per each criterion is highlighted above. As we are already aware that Adjusted R Square is better than R square, we will consider Adjusted R Squared and AIC as comparison criteria. However, both criteria agree on the best model and that is model 5, with 5 parameters i.e. top 5 parameters according to Table 0

8. Variable selection based on automate algorithm.

Step 21. Use the R function "StepAIC" to perform forward variable selection. Compare the result with that in Step 19.

- **Code**

```
??StepAIC

install.packages("MASS")
library(MASS)

?stepAIC
model <- lm(distance ~.,data = faa_std[,-1])

stepAIC(model)
```

- **Relevant Output**

```
Step:  AIC=1916.05
distance ~ no_pasg + speed_air + height + duration + airbus


            Df Sum of Sq       RSS    AIC
- duration   1      9629   3403168 1914.6
<none>                     3393539 1916.0
- no_pasg    1     37010   3430549 1916.2
- height     1   3165734   6559273 2042.6
- airbus     1   8597384  11990923 2160.2
- speed_air  1 125238022 128631561 2622.9
```

- **Observations**

The model suggested by the algorithm is different than the model suggested in earlier steps

- ## Conclusion

According to step 20, the model we finalized was:
distance ~ speed_ground + airbus + height + pitch + duration
However, StepAIC suggests the model:
distance ~ no_pasg + speed_air + height + duration + airbus

Variable	Manual	StepAIC
speed_ground	Yes	No
speed_air	No	Yes
airbus	Yes	Yes
height	Yes	Yes
pitch	Yes	No
duration	Yes	Yes
no_pasg	No	Yes

As we can see above, variables in red were not used by the methods listed, whereas the ones in green were used by either of them and the yellow were used by both methods. As we can see, it chose Speed Air over Speed Ground because of its higher statistical significance, however, what it may not have considered is that it has more missing values and hence may not be useful in accurate prediction. What we eventually conclude is that, different methods for variable selection are meaningful based on our intent or purpose of the model.