BANA 7042

Modeling the Landing Distance using Linear Regression

Name: Ali Aziz

UCID: M12431633

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).

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.

R code:

```
# Step 1: Read in data
library(readxl)
FAA1 <- read_excel("FAA1.xls")
FAA2 <- read_excel("FAA2.xls")
```

Relevant R output:

N/A

Observations:

Both datasets successfully imported.

Conclusion/decision:

Proceed with initial exploration.

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?

R code:

```
# Step 2: Check structure of each data set
str(FAA1)
str(FAA2)
FAA1$aircraft <- factor(FAA1$aircraft)
FAA2$aircraft <- factor(FAA2$aircraft)
```

Relevant R output:

```
> str(FAA1)
Classes 'tbl_df', 'tbl' and 'data.frame':    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 ...
> str(FAA2)
Classes 'tbl_df', 'tbl' and 'data.frame':    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 ...
```

Observations:

The sample size of FAA1 is 800 and it has 8 variables.

The sample size of FAA2 is 150 and it has 7 variables.

Apart from having a much smaller sample size, in the FAA2 dataset the variable "duration" is missing.

Additionally, the variable "aircraft" is being interpreted as a character vector although it is should actually be a factor.

Conclusion/decision:

Changed the "aircraft" variable in both datasets to a factor using the following code:

```
FAA1$aircraft <- factor(FAA1$aircraft)
FAA2$aircraft <- factor(FAA2$aircraft)
```

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?

R code:

```
# Step 3: Merge the two data sets
# Before merging, will add duration column to FAA2
FAA2$duration <- rep(NA, 150)
FAA <- rbind(FAA1, FAA2)

# Check for duplicates
sum(duplicated(FAA[-2]))

# There are 100 duplicates so these should be removed leaving 850 observations
FAA <- FAA[!duplicated(FAA[-2]),]

# Confirm that there are no duplicates
sum(duplicated(FAA[-2])) == 0
```

Relevant R output:

Before removing duplicates:

```
> sum(duplicated(FAA[-2]))
[1] 100
```

After removing duplicates:

```
> sum(duplicated(FAA[-2])) == 0
```

```
[1] TRUE
```

Observations:

There were 100 duplicate observations.

Conclusion/decision:

Removed the duplicate observations.

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.

R code:

```
# Step 4: Check structure of combined data set
str(FAA)
summary(FAA)
apply(FAA[-1], 2, sd, na.rm = TRUE)
```

Relevant R output:

```
> str(FAA)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':          850 obs. of  8 variables:
 $ aircraft    : Factor w/ 2 levels "airbus","boeing": 2 2 2 2 2 2 2 2 2 2 ...
 $ 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 ...
> summary(FAA)
   aircraft      duration         no_pasg      speed_ground     speed_air          height           pitch          distance
 airbus:450   Min.   : 14.76   Min.   :29.0   Min.   : 27.74   Min.   : 90.00   Min.   :-3.546   Min.   :2.284   Min.   :  34.08
 boeing:400   1st Qu.:119.49   1st Qu.:55.0   1st Qu.: 65.90   1st Qu.: 96.25   1st Qu.:23.314   1st Qu.:3.642   1st Qu.: 883.79
              Median :153.95   Median :60.0   Median : 79.64   Median :101.15   Median :30.093   Median :4.008   Median :1258.09
              Mean   :154.01   Mean   :60.1   Mean   : 79.45   Mean   :103.80   Mean   :30.144   Mean   :4.009   Mean   :1526.02
              3rd Qu.:188.91   3rd Qu.:65.0   3rd Qu.: 92.06   3rd Qu.:109.40   3rd Qu.:36.993   3rd Qu.:4.377   3rd Qu.:1936.95
              Max.   :305.62   Max.   :87.0   Max.   :141.22   Max.   :141.72   Max.   :59.946   Max.   :5.927   Max.   :6533.05
              NA's   :50                                       NA's   :642
> apply(FAA[-1], 2, sd, na.rm = TRUE)
    duration      no_pasg speed_ground    speed_air       height        pitch     distance
  49.2592338    7.4931370   19.0594903   10.2590370   10.2877268    0.5288298  928.5600816
```

Observations:

After removing duplicates, the sample size is 850. The number of variables is the same i.e. 8.

Some variables have abnormal values.

There are two aircraft types, there are 450 observations for Airbus flights and 400 observations for Boeing flights.

The variable speed_air has many missing values, 642/850.

The maximum landing distance is 6533 which is greater than the length of a typical runway.

Conclusion/decision:

The abnormal values need to be addressed.

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?

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

1. The sample size after removing duplicates is 850
2. There are 8 variables:
    a. Response variable is landing distance
    b. The 7 predictors are air speed, ground speed, aircraft type, height, pitch, number of passengers and flight duration
3. There were 100 duplicate observations
4. There are two aircraft types Airbus and Boeing with 450 and 400 observations respectively
5. There are two variables with missing values:
    a. Air speed has 642/850 missing values but fortunately high-speed data is available (>90 mph)
    b. Duration has 50/850 missing values

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.

R code:

```
# Step 6:
attach(FAA)

# Take a look at abnormal obervations for each variable
FAA[duration<=40 & is.na(duration) == F,]
FAA[no_pasg <= 0, ] # sanity check
FAA[speed_ground < 30 | speed_ground > 140, ]
FAA[(speed_air < 30 | speed_air > 140) & is.na(speed_air) == F, ]
FAA[height < 6, ]
FAA[distance > 6000, ] # these are not abnormal

# Keep only rows with normal values
FAA <- FAA[(duration > 40 | is.na(duration) == T) & no_pasg > 0 & (speed_ground >= 30 & speed_ground
<= 140) &
   ((speed_air >= 30 & speed_air <= 140) | is.na(speed_air) == T) & height >=6,]

# Check number of remaining rows
nrow(FAA)
```

Relevant R output:

Following variables have abnormal observations:

```
> FAA[duration<=40 & is.na(duration) == F,]
# A tibble: 5 x 8
  aircraft duration no_pasg speed_ground speed_air height pitch distance
  <fctr>      <dbl>   <dbl>        <dbl>     <dbl>  <dbl> <dbl>    <dbl>
1 boeing       17.4    63.0         63.6        NA   28.4  3.94     1032
2 boeing       14.8    59.0          108       109   46.9  4.81     3646
3 boeing       31.4    51.0         98.2      99.1   52.5  4.16     2808
4 airbus       16.9    54.0         94.5      95.9   37.5  4.17     2163
5 airbus       31.7    61.0         76.4        NA   31.0  2.82      948
> FAA[speed_ground < 30 | speed_ground > 140, ]
# A tibble: 3 x 8
  aircraft duration no_pasg speed_ground speed_air height pitch distance
  <fctr>      <dbl>   <dbl>        <dbl>     <dbl>  <dbl> <dbl>    <dbl>
1 boeing        181    54.0          141       142   23.6  5.22     6533
2 boeing        213    61.0         29.2        NA   23.3  4.40     1077
3 boeing        142    46.0         27.7        NA   24.4  4.37     1324
> FAA[(speed_air < 30 | speed_air > 140) & is.na(speed_air) == F, ]
# A tibble: 1 x 8
  aircraft duration no_pasg speed_ground speed_air height pitch distance
  <fctr>      <dbl>   <dbl>        <dbl>     <dbl>  <dbl> <dbl>    <dbl>
1 boeing        181    54.0          141       142   23.6  5.22     6533
> FAA[height < 6, ]
```

```
# A tibble: 10 x 8
   aircraft duration no_pasg speed_ground speed_air  height pitch distance
   <fctr>      <dbl>   <dbl>        <dbl>     <dbl>   <dbl> <dbl>    <dbl>
 1 boeing        284    62.0         58.9        NA  4.26    4.77    426
 2 boeing        175    64.0         52.5        NA -3.55    4.21    581
 3 boeing        146    69.0         71.8        NA -1.53    4.20    739
 4 boeing        133    73.0         57.0        NA  1.25    4.72    371
 5 boeing        124    72.0         60.4        NA  3.79    3.71    642
 6 boeing        120    68.0         70.2        NA  2.21    3.74    816
 7 airbus        103    73.0         93.0        NA -3.33    4.83   1568
 8 airbus        158    68.0         56.5        NA -0.0678  4.69    380
 9 airbus        164    62.0         72.0        NA  0.0861  3.62    538
10 airbus        151    58.0         66.4        NA -2.92    3.12     34.1
```

```
> # Check number of remaining rows
> nrow(FAA)
[1] 832
```

Observations:

There are 18 rows with abnormal values.

- 10 observations with abnormal height values including negative ones
- 5 observations with abnormal durations
- 3 rows with abnormal values for speed ground (one also had abnormal air speed value)

Two flights had landing distances greater than 6000 but these were not removed because these could be correct (long runways).

Conclusion/decision:

Removed the 18 abnormal observations.

Step 7. Repeat Step 4.

R code:

```
# Step 7: Repeat Step 4
str(FAA)
summary(FAA)
apply(FAA[-1], 2, sd, na.rm = T)
```

Relevant R output:

```
> str(FAA)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':         832 obs. of  8 variables:
 $ aircraft    : Factor w/ 2 levels "airbus","boeing": 2 2 2 2 2 2 2 2 2 2 ...
 $ 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 ...
> summary(FAA)
```

```
  aircraft      duration         no_pasg       speed_ground      speed_air          height           pitch       
 airbus:444   Min.   : 41.95   Min.   :29.00   Min.   : 33.57   Min.   : 90.00   Min.   : 6.228   Min.   :2.284  
 boeing:388   1st Qu.:119.65   1st Qu.:55.00   1st Qu.: 66.20   1st Qu.: 96.25   1st Qu.:23.530   1st Qu.:3.640  
              Median :154.26   Median :60.00   Median : 79.83   Median :101.15   Median :30.185   Median :4.002  
              Mean   :154.73   Mean   :60.06   Mean   : 79.61   Mean   :103.65   Mean   :30.474   Mean   :4.005  
              3rd Qu.:189.64   3rd Qu.:65.00   3rd Qu.: 91.99   3rd Qu.:109.40   3rd Qu.:37.018   3rd Qu.:4.370  
              Max.   :305.62   Max.   :87.00   Max.   :136.66   Max.   :136.42   Max.   :59.946   Max.   :5.927  
              NA's   :50                                        NA's   :628                                      
    distance     
 Min.   :  41.72  
 1st Qu.: 893.43  
 Median :1263.54  
 Mean   :1528.24  
 3rd Qu.:1937.58  
 Max.   :6309.95  

> apply(FAA[-1], 2, sd, na.rm = T)
    duration       no_pasg speed_ground    speed_air       height        pitch     distance 
  48.3350296    7.4880567   18.8288115    9.9823067    9.7906666    0.5262829  911.0450647 
```

Observations:

The sample size is now 832 and there are 8 variables.

Conclusion/decision:

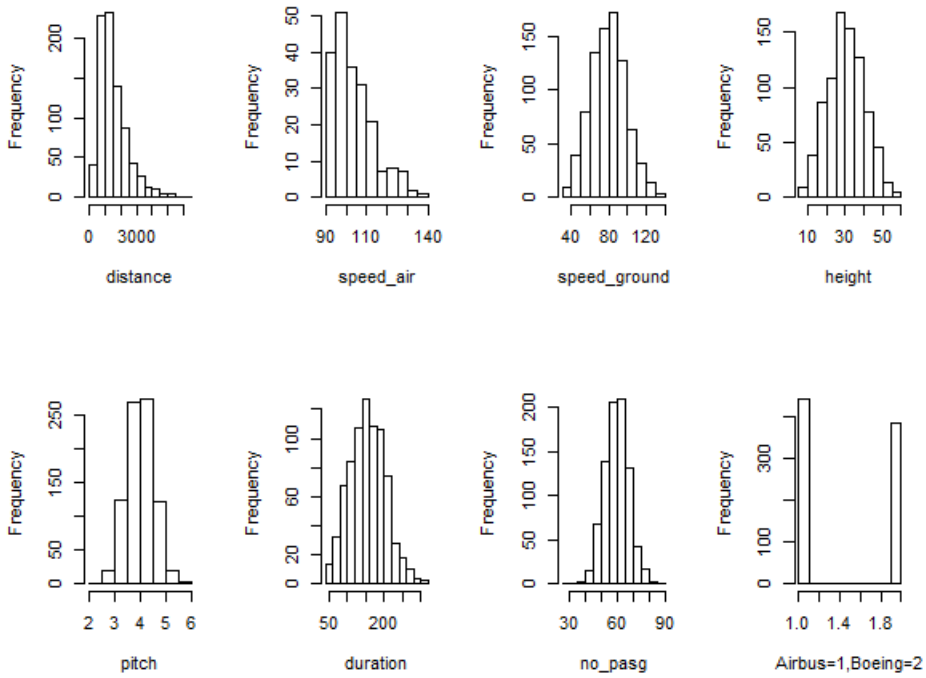
Now that data is clean, can proceed with further exploration.

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

R code:

```
# Step 8: Draw histograms for each variable
par(mfrow = c(2,4))
hist(distance, main = "")
hist(speed_air, main = "")
hist(speed_ground, main = "")
hist(height, main = "")
hist(pitch, main = "")
hist(duration, main = "")
hist(no_pasg, main = "")
hist(as.numeric(aircraft), main = "", xlab = "Airbus=1,Boeing=2")
```

Relevant R output:



Observations:

Most of the continuous variables are approximately normally distributed.

The exceptions are distance and air speed.

Air speed distribution looks like a truncated normal due to missing values for speeds below 90 mph.

The bar chart for aircraft type shows that there are more observations for Airbus flights than for Boeing ones.

Conclusion/decision:

The information gained from these histograms should be kept in mind, especially the fact that landing distance does not look normally distributed.

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

1. The sample size for the 8 variables after removing 18 observations with abnormal values is 832
2. Two flights had landing distances greater than 6000 but these were not removed because these could be correct (long runways)
3. The 18 observations with abnormal values consisted of:
    a. 10 observations with abnormal height values including negative ones
    b. 5 observations with abnormal durations

    c. 3 rows with abnormal values for speed ground (one also had abnormal air speed value)
4. Variables are approximately normally distributed except for distance and air speed
5. Air speed distribution looks like a truncated normal due to missing values for speeds less than 90 mph

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.

R code:

```
# Step 10: Compute pairwise correlation
cor(FAA[,2:7], FAA[,8], use = "complete.obs")
cor(as.numeric(aircraft), distance, use = "complete.obs")
```

Relevant R output:

```
> cor(FAA[,2:7], FAA[,8], use = "complete.obs")
                distance
duration      0.03671284
no_pasg      -0.01879416
speed_ground  0.93171687
speed_air     0.94529684
height        0.08568026
pitch         0.03723416
> cor(as.numeric(aircraft), distance, use = "complete.obs")
[1] 0.2375433
```

Observations:

Table 1. Correlation between landing distance and each predictor

Predictor	**Size of correlation**	**Direction of correlation**
Air speed	0.945	Positive
Ground speed	0.932	Positive
Aircraft type	0.238	Positive (going from Airbus to Boeing)
Height	0.086	Positive
Duration	0.055	Negative
Pitch	0.037	Positive
Number of passengers	0.019	Negative

In Table 1 it can be seen that:

- Landing distance is strongly and positively correlated with air speed and ground speed
- There is weak positive correlation between distance and aircraft type suggesting that Boeing flights have larger landing distances
- Landing distance has very weak correlation, if any, with height, pitch, duration and number of passengers

Conclusion/decision:

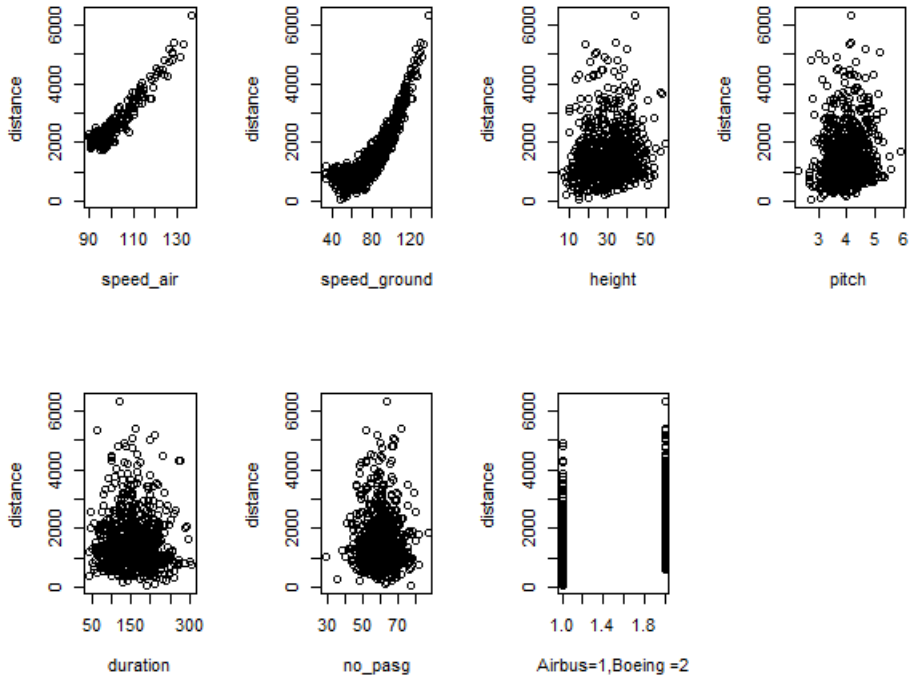
Based on the correlation values, air speed, ground speed and aircraft could be the most important factors that impact the landing distance.

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?

R code:

```
# Step 11: Scatter plots
plot(speed_air, distance)
plot(speed_ground, distance)
plot(height, distance)
plot(pitch, distance)
plot(duration, distance)
plot(no_pasg, distance)
plot(as.numeric(aircraft), distance, xlab = "Airbus=1,Boeing =2")
```

Relevant R output:



Observations:

The correlation strength observed in the plots is consistent with the values computed in Step 10.

A strong, increasing trend in landing distance can be seen in the scatter plots of distance against ground speed and air speed.

A weak, increasing trend in landing distance can be observed when going from Airbus flights to Boeing flights.

The scatter plots of the rest of the variables do not seem to have any distinct trend.

Conclusion/decision:

The scatter plots give us more confidence about the conclusions we reached in Step 10 i.e. the variables air speed, ground speed and aircraft could be the most important factors that impact the landing distance.

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.

Yes, airplane make was included in Steps 10-11 (Airbus coded as 1, Boeing coded as 2).

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.

R code:

```
# Step 13: Regress Y (landing distance) on each of the X variables
summary(lm(distance ~ speed_air))
summary(lm(distance ~ speed_ground))
summary(lm(distance ~ height))
summary(lm(distance ~ pitch))
summary(lm(distance ~ duration))
summary(lm(distance ~ no_pasg))
summary(lm(distance ~ aircraft))
```

Relevant R output:

The size of the p-value and the direction of the regression coefficient for each variable is given in Table 2 below.

Table 2. Predictors ranked based on significance

Predictor	**P-value**	**Direction of regression coefficient**
Air speed	<2e-16	Positive
Ground speed	<2e-16	Positive
Aircraft type	2.28e-12	Positive (going from Airbus to Boeing)
Height	6.75e-05	Positive
Pitch	0.00272	Positive
Duration	0.0792	Negative
Number of passengers	0.377	Negative

Observations:

As expected air speed, ground speed and aircraft type have the 3 lowest p-values.

The variable height also has a very low p-value.

If we take the level of significance to be 0.01 then pitch is also significant.

Conclusion/decision:

Air speed, ground speed, aircraft type are the 5 significant variables.

The variables duration and number of passengers are not significant.

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.

R code:

```
# Step 14: Standardize each X variable to X' and then regress Y (landing distance) on each of the X' variables

scaled.FAA <- data.frame(scale(FAA[-1]))
summary(scaled.FAA)
apply(scaled.FAA, 2, sd, na.rm = T)
attach(scaled.FAA)
summary(lm(distance ~ speed_air))
summary(lm(distance ~ speed_ground))
summary(lm(distance ~ height))
summary(lm(distance ~ pitch))
summary(lm(distance ~ duration))
summary(lm(distance ~ no_pasg))
attach(FAA)
```

Relevant R output:

The size of the regression coefficient and the direction of the regression coefficient for each variable after normalization is given in Table 3 below.

Table 3. Predictors ranked based on size of regression coefficient after normalization of the variables.

Predictor	**Size of regression coefficient**	**Direction of regression coefficient**
Aircraft type	219.340	Positive (going from Airbus to Boeing)
Air speed	0.888	Positive
Ground speed	0.866	Positive
Height	0.107	Positive
Pitch	0.0875	Positive
Duration	0.0558	Negative
Number of passengers	0.0141	Negative

Observations:

The order of the variables based on the size of the regression coefficient after normalization is the same except that aircraft type moves from third place to first place.

Conclusion/decision:

The results are consistent. Although aircraft type is now at the top, this can be explained by the fact that it is a factor variable with just 2 levels, so normalization does not carry the same weight as it does with the other continuous variables.

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.

The results are consistent barring a couple of exceptions mentioned above:

1) Duration has a negative regression coefficient but positive correlation with landing distance. However,the magnitude of the correlation is very small, the variable and is not significant, and the ranking of the variable is the same in all the tables.

2) After normalization, aircraft type is ranked first in terms of the size of the regression coefficient. But this is not a fair comparison because aircraft type is a factor variable whereas the other variables are continuous. Based on Tables 1 and 2, it should remain at the number 3 spot.

Table 0. Predictors ranked based on their relative importance in determining the landing distance

Predictor	**Ranking**
Air speed	1
Ground speed	2
Aircraft type	3
Height	4
Pitch	5
Duration	6
Number of passengers	7

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?

R code:

```
# Step 16: Compare the regression coefficients of the three provided models
summary(lm(distance ~ speed_ground))
summary(lm(distance ~ speed_air))
summary(lm(distance ~ speed_ground + speed_air))

cor(speed_ground, speed_air, use = "complete.obs")
```

Relevant R output:

```
> summary(lm(distance ~ speed_ground))


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1808.3445    68.6579  -26.34   <2e-16 ***
speed_ground    41.9109     0.8393   49.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 455.5 on 830 degrees of freedom
Multiple R-squared:  0.7503,   Adjusted R-squared:   0.75
F-statistic:  2494 on 1 and 830 DF,  p-value: < 2.2e-16

> summary(lm(distance ~ speed_air))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5604.993    206.915  -27.09   <2e-16 ***
speed_air      81.016      1.987   40.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 282.6 on 202 degrees of freedom
  (628 observations deleted due to missingness)
Multiple R-squared:  0.8916,   Adjusted R-squared:  0.8911
F-statistic:  1662 on 1 and 202 DF,  p-value: < 2.2e-16

> summary(lm(distance ~ speed_ground + speed_air))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  -5611.84     206.93 -27.120  < 2e-16 ***
speed_ground   -14.03      12.98  -1.081    0.281
speed_air       95.10      13.18   7.216 1.07e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 282.5 on 201 degrees of freedom
  (628 observations deleted due to missingness)
Multiple R-squared:  0.8923,   Adjusted R-squared:  0.8912
F-statistic: 832.3 on 2 and 201 DF,  p-value: < 2.2e-16

> cor(speed_ground, speed_air, use = "complete.obs")
[1] 0.9885771
```

Observations:

There is a significant change. The sign of the regression coefficient for speed ground goes from positive to negative when speed air is also included in the model.

This is explained by the very high correlation between the two speeds, the correlation coefficient is 0.989.

Conclusion/decision:

Although air speed has many missing values, I would keep air speed in the model in the model and drop ground speed for the following reasons:

1) The missing values of air speed are only for speeds below 90 mph. The goal of the project is to predict landing overrun and all the values for the high-speed region are present.
2) Air speed gives a better R-squared value, has slightly stronger correlation and a slightly larger regression coefficient compared to ground speed.
3) Based on the scatter plots, the relationship between ground speed and landing distance is actually quadratically shaped. So trying to fit a single linear model over all ground speeds actually has an adverse impact on prediction accuracy in the crucial high-speed region.

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?

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

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

R code:

```
# Steps 17, 18 and 19: Fit 6 models according to variable ranking and plot 3 model selection criteria

r.squared <- rep(0,6)
adjusted.r.squared <- rep(0,6)
aic <- rep(0,6)

# Model 1: LD ~ Air speed
model <- lm(distance ~ speed_air)
r.squared[1] <- summary(model)$r.squared
adjusted.r.squared[1] <- summary(model)$adj.r.squared
aic[1] <- AIC(model)

# Model 2: LD ~ Air speed + Ground speed
model <- lm(distance ~ speed_air + speed_ground)
r.squared[2] <- summary(model)$r.squared
adjusted.r.squared[2] <- summary(model)$adj.r.squared
aic[2] <- AIC(model)

# Model 3: LD ~ Air speed + Ground speed + Aircraft type
model <- lm(distance ~ speed_air + speed_ground + aircraft)
r.squared[3] <- summary(model)$r.squared
adjusted.r.squared[3] <- summary(model)$adj.r.squared
aic[3] <- AIC(model)

# Model 4: LD ~ Air speed + Ground speed + Aircraft type + height
```

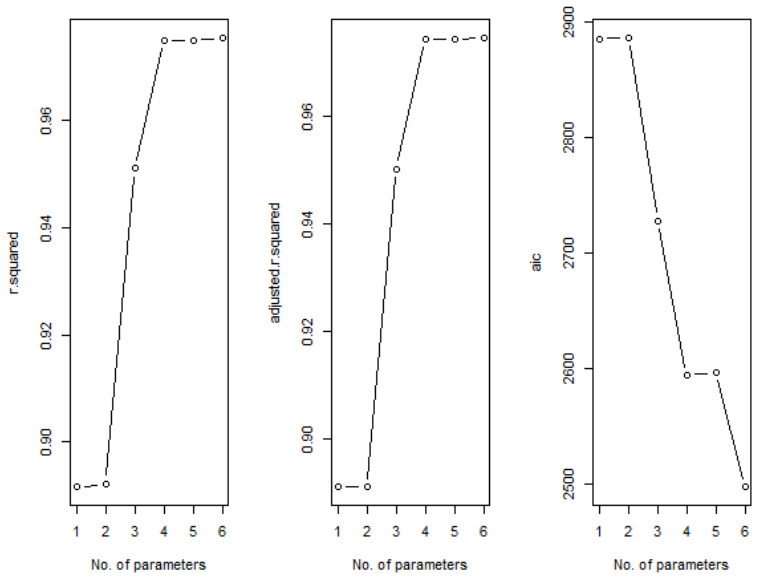
```
model <- lm(distance ~ speed_air + speed_ground + aircraft + height)
r.squared[4] <- summary(model)$r.squared
adjusted.r.squared[4] <- summary(model)$adj.r.squared
aic[4] <- AIC(model)

# Model 5: LD ~ Air speed + Ground speed + Aircraft type + height + pitch
model <- lm(distance ~ speed_air + speed_ground + aircraft + height + pitch)
r.squared[5] <- summary(model)$r.squared
adjusted.r.squared[5] <- summary(model)$adj.r.squared
aic[5] <- AIC(model)

# Model 6:  LD ~ Air speed + Ground speed + Aircraft type + height + duration
model <- lm(distance ~ speed_air + speed_ground + aircraft + height + pitch + duration)
r.squared[6] <- summary(model)$r.squared
adjusted.r.squared[6] <- summary(model)$adj.r.squared
aic[6] <- AIC(model)

# Plots
par(mfrow = c(1,3))
# R squared vs number of parameters
plot(1:6, r.squared, type = "b", xlab = "No. of parameters")
# Adjusted R squared vs number of parameters
plot(1:6, adjusted.r.squared, type = "b", xlab = "No. of parameters")
# AIC vs number of parameters
plot(1:6, aic, type = "b", xlab = "No. of parameters")
```

Relevant R output:



Observations:

Since R-squared does not take into account model complexity (number of parameters), it always increases when new parameters are added.

When calculating adjusted R squared, a penalty is paid for the number of parameters. We see that it decreases slightly when the variable pitch is added. However, it is still highest when all variables are present.

Similarly, AIC also penalizes parameter additions. Again, we see a slight increase in AIC (for AIC lower is better) when pitch is added to the model. But the lowest AIC value is obtained when all 6 predictors are in the model.

Conclusion/decision:

R squared is not a good criterion for model comparisons because model complexity is not considered.

Based solely on the adjusted R squared and AIC values, the model with the top 6 predictors seems to be the best option.

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

Although the best values of adjusted R squared and AIC values are obtained when the top 6 predictors are used, to build a predictive model I would just pick the 3 variables air speed, aircraft and height. Adding both pitch and duration does improve the in-sample prediction of the model, but for out-of-sample prediction there is no guarantee that this will still hold. Furthermore, as mentioned earlier, due to the very high correlation between air speed and ground speed it is enough to keep air speed.

By selecting fewer variables the aim is to prevent “overfitting” and to obtain a model that would give good predictions out-of-sample. Furthermore, having fewer variables means that we can avoid the multi-collinearity problem (between air speed and ground speed for example), we can better interpret our results and we require less data for making predictions.

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.

R code:

#Step 21: Use the R function "StepAIC" to perform forward variable selection. Compare the result with that in Step 19.
library(MASS)
fit1 <- lm(distance ~ ., na.omit(FAA))
fit2 <- lm(distance ~ 1, na.omit(FAA))
stepAIC(fit1,direction="backward")$anova
stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))$anova
stepAIC(fit2,direction="both",scope=list(upper=fit1,lower=fit2))$anova

Relevant R output:

Only showing partial output for forward direction. Final model for all 3 directions was the same.

```
> stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
distance ~ 1

Final Model:
distance ~ speed_air + aircraft + height


         Step Df  Deviance Resid. Df Resid. Dev      AIC
1                                195  146041225 2652.172
2 + speed_air  1 130500410       194   15540815 2215.050
3  + aircraft  1   8421618       193    7119197 2064.037
4    + height  1   3479278       192    3639920 1934.554
```

Observations:

The final model chosen by stepAIC has the variables air speed, aircraft and height.

In step 19, the approach was to add the next most important variable which had the most impact when taken individually and then to see if the AIC decreased or not. So if one were to stop adding variables when AIC did not decrease or if one were to pick the combination that gave the lowest AIC, then either way the resulting model would be different from the model returned by the stepAIC approach. This is because the stepAIC algorithm checks the additional value each variable will bring and then selects the best one (the one that gives the lowest AIC).

Conclusion/decision:

The final model chosen by stepAIC is the better choice here as it chooses the next variable to add to the model by picking the one that brings the most additional value (the concept of putting together a

“team” of variables that complement each other). So the final model would use air speed, aircraft type and height to predict landing distance. This also matches the model I selected in Step 20 based on the AIC/adjusted R squared values of the 6 models and my knowledge of the variables.