BANA 7042 STATISTICAL MODELING - PROJECT PART - 2

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 using ***Logistic Regression***

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.

Cleaned Data set

- **R code**

```r
# create cleaned data set
#import data
library(readxl)
library(tidyverse)
library(magrittr)
FAA1 <- read_excel("FAA1.xls")
FAA2 <- read_excel("FAA2.xls")
#remove duplicates if any within data set
FAA1 <- unique(FAA1)
FAA2 <- unique(FAA2)
#finally create unique dataset
merged <- rbind(FAA1[,-2],FAA2)
merged <- unique(merged)
FAA <- merge(merged,FAA1,by=names(merged), all.x=TRUE)
#removing abnormal values
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 )
```

- **Conclusion**

The final cleaned data set has 8 variables and 831 observations

Step 1. Create binary responses

From now on, please work on the cleaned FAA data set you prepared by carrying out Steps 1-9 in Part 1 of the project. Create two binary variables below and attach them to your data set.

- long.landing = 1 if distance > 2500; =0 otherwise
- risky.landing = 1 if distance > 3000; =0 otherwise.

Discard the continuous data you have for “distance”, and assume we are given the binary data of “long.landing” and “risky.landing” only.

- ***R code***

```
#step 1 create binary variables
faa <- faa_clean
faa$long.landing <- 0
faa$risky.landing <- 0

faa[which(faa$distance > 2500),"long.landing"] <- 1
faa[which(faa$distance > 3000),"risky.landing"] <- 1

nrow(faa[which(faa$distance > 2500),])
sum(faa$long.landing)
nrow(faa[which(faa$distance > 3000),])
sum(faa$risky.landing)

t(names(faa))
faa <- faa[,-7]
```

- ***Observations***

We have 61 records classified as Risky and 103 that are classified as long

- ***Conclusion***

The final data set now has 831 observations and 9 variables. The distance variable has been removed:

Variable	Data type	Sample Values
aircraft	chr	airbus "airbus" "airbus" "airbus"
no_pasg	num	36 38 40 41 43 44 45 45 45 45
speed_ground	num	47.5 85.2 80.6 97.6 82.5
speed_air	num	NA NA NA 97 NA
height	num	14 37 28.6 38.4 30.1
pitch	num	4.3 4.12 3.62 3.53 4.09

duration	num	172 188 93.5 123.3 109.2
long.landing	num	0 1 0 1 0 1 0 0 0 0
risky.landing	num	0 0 0 0 0 1 0 1 1 0

Table 1- Cleaned dataset variables

Step 2. Distribution of Long Landing

Use a pie chart or a histogram to show the distribution of "long.landing".

- ***R code***

```
#step 2
library(ggplot2)
ggplot(data = faa, aes(x = as.factor(long.landing))) +
geom_histogram(stat =    "count")
```

- ***Code Output***

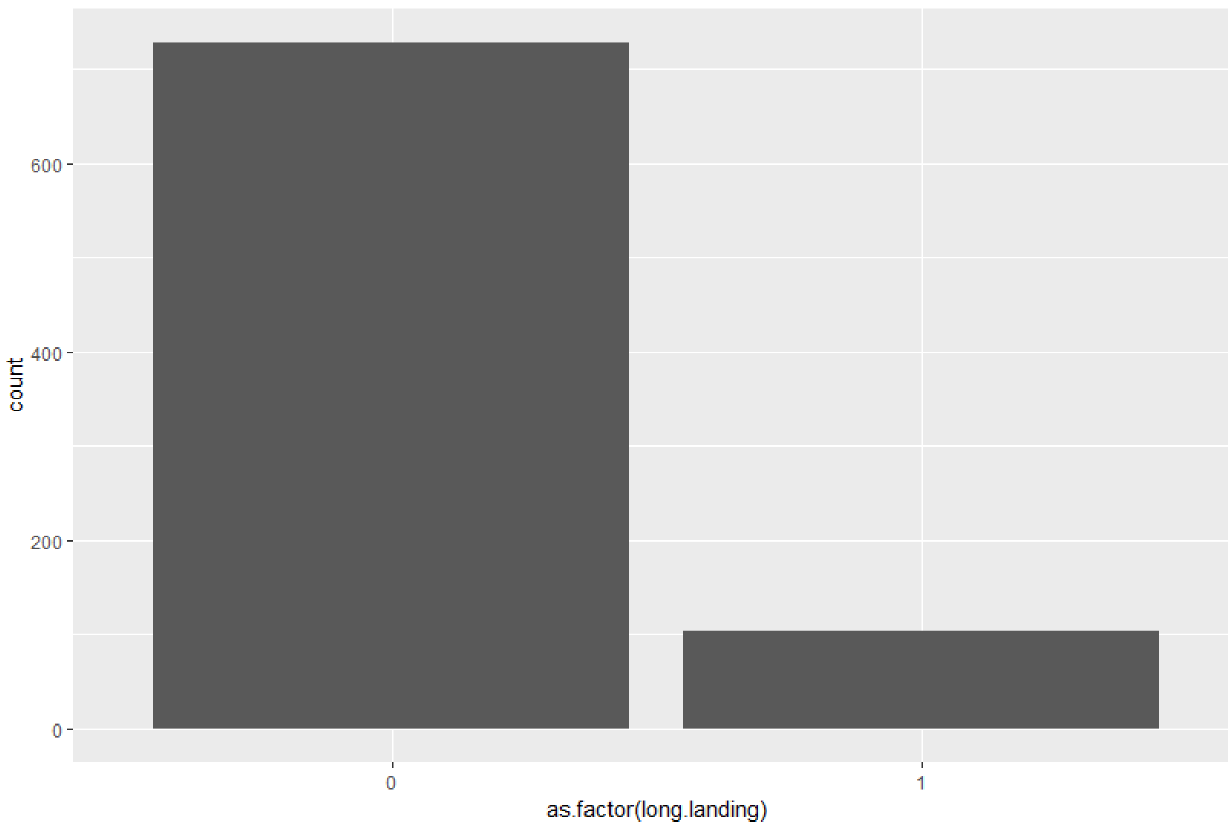

Figure 1 - Distribution of Long Distance

- ***Observations***

We observe that there are more 0s than there are 1s.

Value	Count	Percent

0	728	87.60529
1	103	12.39471

Table 2 - Count by values for Long Distance

- ***Conclusion***

Mostly, the landings are not long.

Step 3. Single-factor regression analysis

Perform single-factor regression analysis for each of the potential risk factors, in a similar way to what you did in Steps 13-15 of Part 1. But here the response “long.landing” is binary. You may consider using logistic regression.

Provide a table that ranks the factors from the most important to the least. This table contains 5 columns: the names of variables, the size of the regression coefficient, the odds ratio, the direction of the regression coefficient (positive or negative), and the p-value.

- ***R code***

```
#step 3
t(names(faa))
faa$aircraft <- as.factor(faa$aircraft)
var_name <- rep('',7)
coeff <- rep(0,7)
odds_ratio <- rep(0,7)
direction <- rep('+',7)
p_val <- rep(0,7)
j <- 1

for(i in c(1,2,3,4,5,6,7)) {
  fit <- glm(long.landing ~ faa[,i],family=binomial(link='logit'),data=faa)
  var_name[j] <- names(faa)[i]
  coeff[j] <- abs(summary(fit)$coefficients[2,1])
  odds_ratio[j] <- exp(fit$coefficients[2])
    if(summary(fit)$coefficients[2,1] < 0) {direction[j] <- '-'}
  p_val[j] <- summary(fit)$coefficients[2,4]
  j <- j+1
}
tt <- cbind(1:7,var_name,coeff,odds_ratio,direction,p_val)
```

- ***Code Output***

Sr	Variable	Coeff	Odds Ratio	Direction	P value	Rank
3	speed_ground	0.472345752	1.603751789	+	3.93534E-14	1
4	speed_air	0.512321766	1.669162103	+	4.33412E-11	2
1	Aircraft = Boeing	0.86411986	2.372916667	+	8.39859E-05	3
6	pitch	0.400527824	1.492612326	+	0.046649818	4
5	height	0.008623997	1.008661291	+	0.42185757	5
2	no_pasg	0.007256406	0.992769858	-	0.605856519	6

7	duration	0.001070492	0.998930081	-	0.630512185	7

Table 3 - Variable Ranks on P value

- ***Observations***

The ranks are calculated using p-values based on the criteria for ranking in Project Part 1 step 13-15

- ***Conclusion***

We observe that Speed Ground is the most significant variable vs Duration which seems to be the least based on p-values on individual logistic regressions.
Based on 95% level of significance, we identify Speed Ground, Speed Air, Aircraft and Pitch to be statistically significant. For the binary predictor Aircraft, the odds of a long landing are more when the aircraft is Boeing

Step 4. Visualize Association

For those significant factors identified in Step 3, visualize its association with “long.landing”. See the slides (pp. 12-21) for Lecture 3.

- ***R code***

```
#step 4
library(car)
names(faa)
scatterplotMatrix(~long.landing + no_pasg + speed_ground +
                    speed_air + height + pitch + duration, data <- faa,
                  regLine = F, ellipse = F, diagonal = F,smooth = F  )

par(mfrow = c(1,3))
plot(long.landing~speed_ground,data <- faa)
plot(long.landing~speed_air,data <- faa)
plot(long.landing~pitch,data <- faa)

par(mfrow = c(2,2))
plot( jitter(long.landing,0.1)~jitter(speed_ground),data <- faa,
      xlab = 'Speed Ground',ylab = 'Long Landing')
plot( jitter(long.landing,0.1)~jitter(speed_air),data <- faa,
      xlab = 'Speed Air',ylab = 'Long Landing')
plot( jitter(long.landing,0.1)~jitter(pitch),data <- faa,
      xlab = 'Pitch',ylab = 'Long Landing')
plot( jitter(long.landing,0.1)~jitter(as.numeric(aircraft)),data <- faa,
      xlab = 'Aircraft',ylab = 'Long Landing')

install.packages("ggpubr")
library(ggpubr)
g_ground <- ggplot(data <- faa,aes(x=speed_ground,fill=factor(long.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
       colour=factor(long.landing)),alpha = 0.5)

g_air <- ggplot(data <- faa,aes(x=speed_air,fill=factor(long.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
   colour=factor(long.landing)),alpha = 0.5)
```

```
g_pitch <- ggplot(data <- faa,aes(x=pitch,fill=factor(long.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
   colour=factor(long.landing)),alpha = 0.5)

g_aircraft <- ggplot(data <- faa,aes(x=aircraft,fill=factor(long.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
  colour=factor(long.landing)),alpha = 0.5)

ggarrange(g_ground,g_air,g_pitch,g_aircraft, ncol = 2, nrow = 2)
```

- ***Code Output***

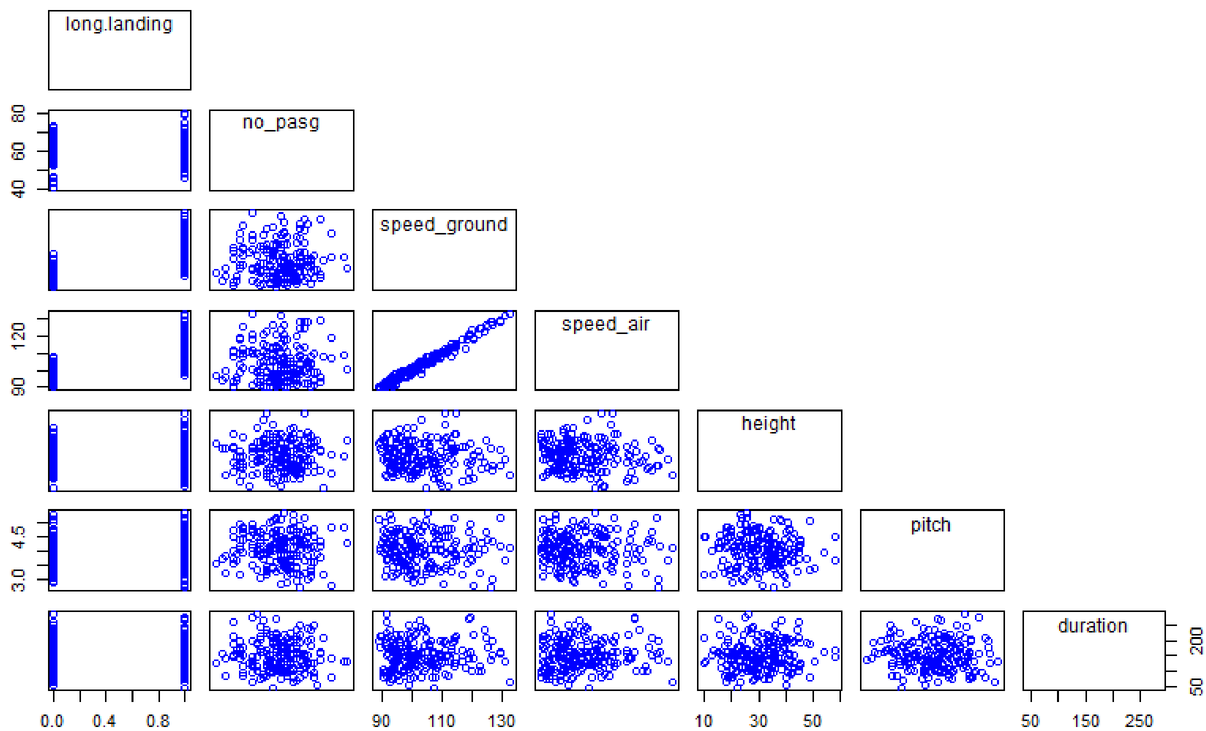

Figure 2 Scatter plot matrix for all variables

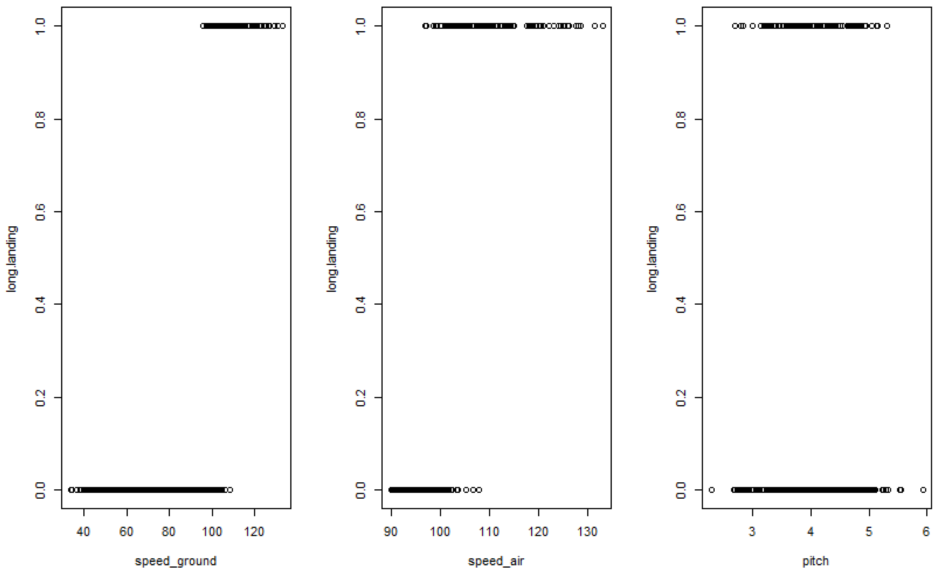

Figure 3 Scatter plot for Significant variables

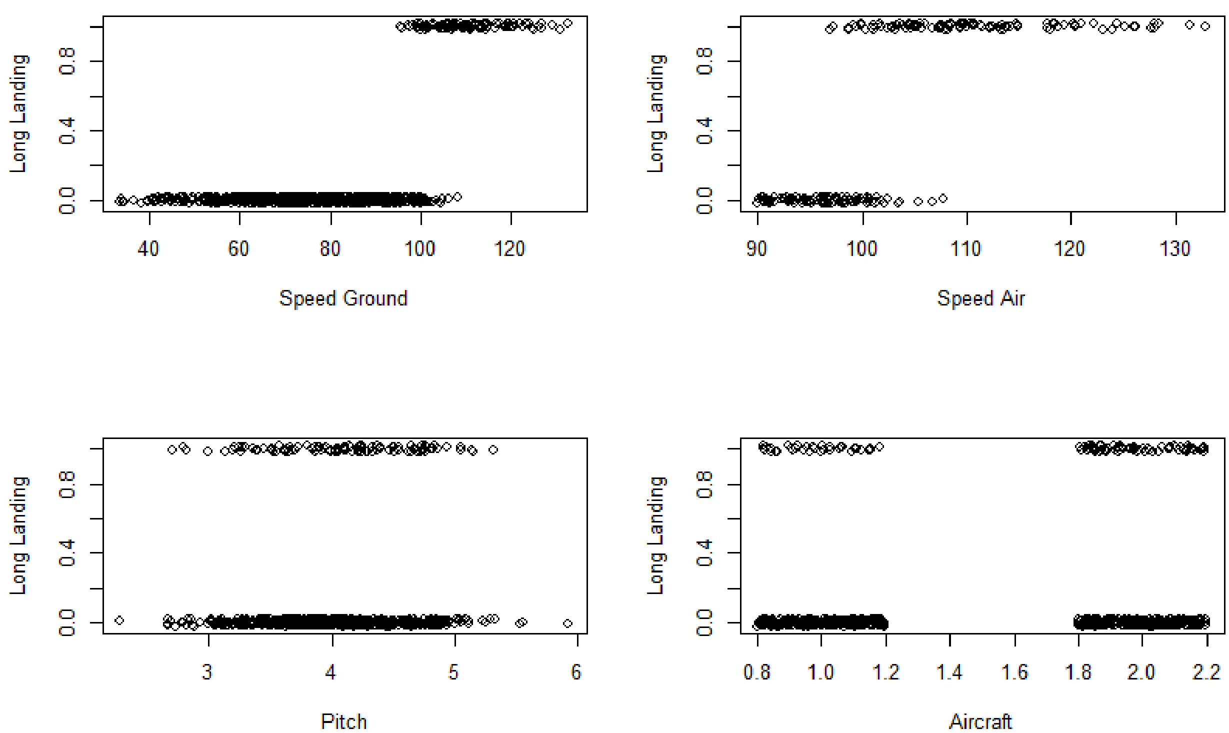

Figure 4 Jitter plot for Significant variables including Aircraft

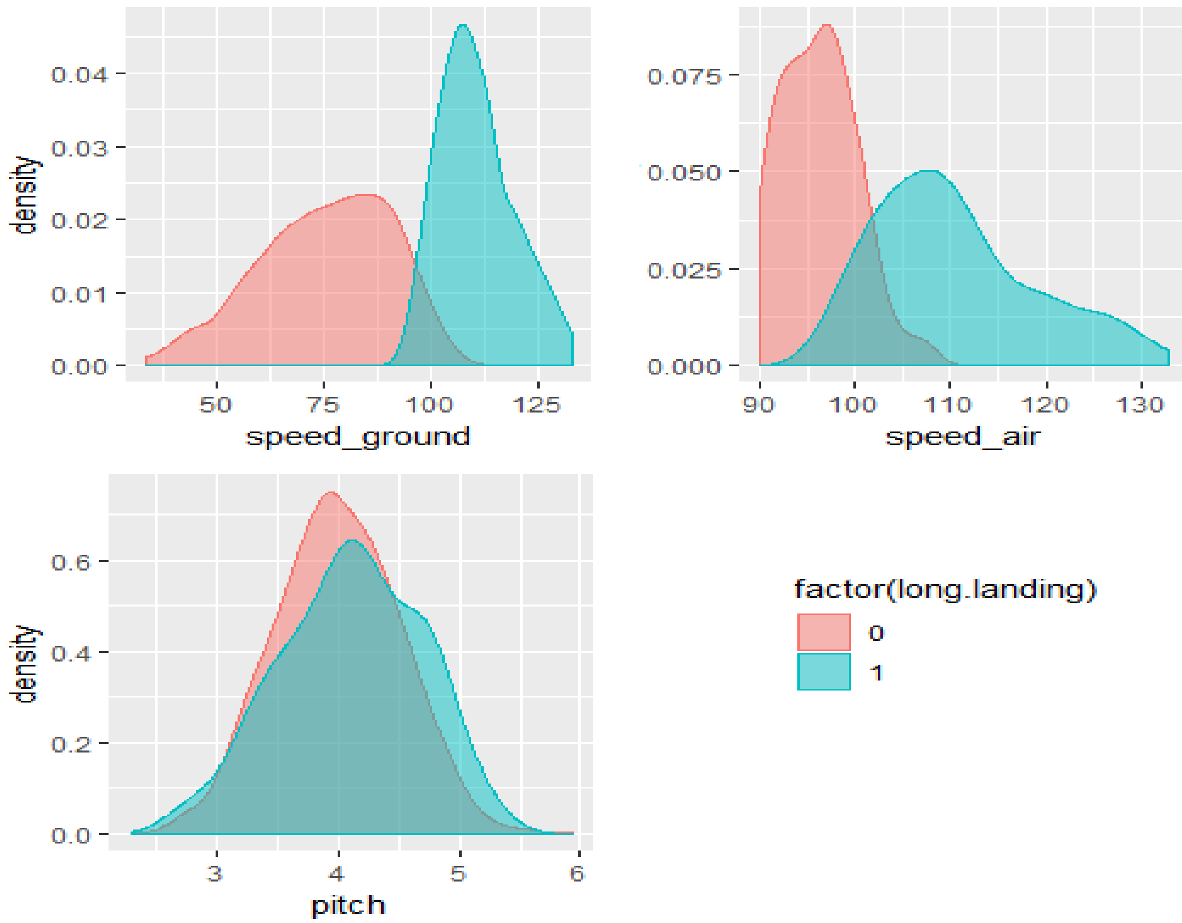

Figure 5 Density plot for Significant variables

- ***Observations***

We visualize the association between Long Landing and all other significant variables using different techniques in R

- ***Conclusion***

Based on the observations of the graphs, we conclude that

- Pitch – The distribution for Long Landing does not seem discriminatory enough
- Speed Ground – A clear distinction between can be made that when Speed Ground is more than 90 mph, it is likely that the landing will be long
- Speed Air - A distinction between can be made that when Speed Ground is more than 95 mph, it is likely that the landing will be long
- Aircraft – We can see that the Landing is more likely to be long when the aircraft is Boeing versus when it is Airbus

Step 5. Model Fitting

Based on the analysis results in Steps 3-4 and the collinearity result seen in Step 16 of Part 1, initiate a “full” model. Fit your model to the data and present your result.

- ***R code***

```
#step 5
fit <- glm(long.landing ~ aircraft + speed_ground + pitch,
           family=binomial(link='logit'),data=faa)
summary(fit)
```

- ***Code Output***

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.11589  -0.01116  -0.00026   0.00000   2.40741

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -67.92855   10.48408  -6.479 9.22e-11 ***
aircraftboeing   3.04348    0.73345   4.150 3.33e-05 ***
speed_ground     0.61471    0.09184   6.694 2.18e-11 ***
pitch            1.06599    0.60389   1.765   0.0775 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 622.778  on 830  degrees of freedom
Residual deviance:  81.309  on 827  degrees of freedom
AIC: 89.309

Number of Fisher Scoring iterations: 10
```

- ***Observations***

We observe that the variable Pitch which was significant earlier is now not significant at level of significance 95%

- ***Conclusion***

Since Speed Air has missing values, based on our observations in Part 1 Step 16, we have not included it in the full model presented above since Speed Air and Speed Ground is highly significant. Also, Pitch, which was significant as such, seem to lose significance in a broader context when other variables are included

Step 6. Forward Step - AIC

Use the R function "Step" to perform forward variable selection using AIC. Compare the result with the table obtained in Step 3. Are the results consistent?

- ***R code***

```
#step 6
null.model <- glm(long.landing ~ 1 , family=binomial(link='logit'),data=faa)
full.model <- glm(long.landing ~ aircraft + no_pasg + speed_ground + height +
                    pitch+ duration ,family=binomial(link='logit'),data=faa)
AIC.model <- step(null.model, scope=list(lower=null.model, upper=full.model),
                     direction='forward')
summary(AIC.model)
```

- ***Code Output***

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.20284  -0.00054   0.00000   0.00000   2.35719

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -119.77598   24.41821  -4.905 9.33e-07 ***
speed_ground      1.02266    0.20290   5.040 4.65e-07 ***
aircraftboeing    5.13443    1.18091   4.348 1.37e-05 ***
height            0.25795    0.06861   3.760  0.00017 ***
pitch             1.53751    0.84109   1.828  0.06755 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 622.778  on 830  degrees of freedom
Residual deviance:  53.204  on 826  degrees of freedom
AIC: 63.204

Number of Fisher Scoring iterations: 12
```

- ***Observations***

The Forward Step selection selects Height as a significant variable and simultaneously Pitch, although is part of the final iteration, is insignificant

- ***Conclusion***

Compared to the model based on inferences drawn from Step 3-4, Height was not a significant variable. However, when we perform Step-Forward-AIC model selection, it does select Height as significant. Furthermore, Pitch is observed as insignificant which was rather significant when we considered it individually. So we conclude that the results are not entirely consistent with our model in step 3. This also brings us to an important idea that when there is less information at hand (less rows or les variables), the model tries its best to explain the variability in response. However, only when we augment the information with more predictors and rows, the model is able to call better decisions.

Step 7. Forward Step - BIC

Use the R function "Step" to perform forward variable selection using BIC.

Compare the result with that from the previous step.

- ***R code***

```
BIC.model <- step(null.model, scope=list(lower=null.model, upper=full.model),
                  direction='forward',k=log(nrow(faa)))
summary(BIC.model)
```

- ***Code Output***

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.43442  -0.00117   0.00000   0.00000   2.57435

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -102.95437   19.22882  -5.354 8.59e-08 ***
speed_ground      0.92657    0.17242   5.374 7.70e-08 ***
aircraftboeing    5.04813    1.11520   4.527 5.99e-06 ***
height            0.23106    0.05959   3.877 0.000106 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 622.778  on 830  degrees of freedom
Residual deviance:  57.047  on 827  degrees of freedom
AIC: 65.047

Number of Fisher Scoring iterations: 11
```

- ***Observations***

We observe that Pitch is no more included in the model. Although the direction of impact of the variables remain the same, their size is different from earlier. IN the following table we compare the results of AIC and BIC model

- ***Conclusion***

Properties	Step Forward AIC					Step Forward BIC				
Deviance Residuals	Min	1Q	Median	3Q	Max	Min	1Q	Median	3Q	Max
	-2.20284	-0.00054	0	0	2.3572	-2.43442	-0.00117	0	0	2.5744
Coefficients	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**		**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	-119.776	24.41821	-4.905	**0.000000933**		-102.954	19.22882	-5.354	**8.59E-08**	
speed_ground	1.02266	0.2029	5.04	**0.000000465**		0.92657	0.17242	5.374	**0.000000077**	
aircraftboeing	5.13443	1.18091	4.348	**0.0000137**		5.04813	1.1152	4.527	**0.00000599**	
height	0.25795	0.06861	3.76	**0.00017**		0.23106	0.05959	3.877	**0.000106**	
pitch	1.53751	0.84109	1.828	**0.06755**		NA				
Null deviance on degrees of freedom	622.778 on 830					622.778 on 830				
Residual deviance on degrees of freedom	53.204 on 826					57.047 on 827				
AIC	63.204					65.047				
BIC	86.81731					83.9372				
Number of Fisher Scoring iterations	12					11				

Table 4 AIC vs BIC model selection

We conclude that the BIC model does not select pitch in the final model. The AIC model did determine Pitch's insignificance, however, kept it in the model. There are also slight variations in the effect of the predictors. We choose the BIC model over AIC model because of reduced model complexity.

Step 8. Intermediate Summary of findings for Long Distance

You are scheduled to meet with an FAA agent who wants to know "what are risk factors for long landings and how do they influence its occurrence?". For your presentation, you are only allowed to show. The question is: what model/ table/ figures/ statements you would include in your presentation. Be selective! One model, One table, No more than three figures, & No more than five bullet statements. Please use statements that she can understand.

- ***R code***

```
#step 8
g_craft <- ggplot(data <- faa,aes(x=factor(aircraft),y=factor(risky.landing)))+
  geom_jitter(position="jitter",aes(colour=factor(risky.landing)),alpha = 0.5);

g_height <- ggplot(data <- faa,aes(x=height,y=factor(risky.landing)))+
  geom_jitter(position="jitter",binwidth=5,aes(
    colour=factor(risky.landing)),alpha = 0.5);

g_ground <- ggplot(data <- faa,aes(x=speed_ground,fill=factor(risky.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                                              colour=factor(risky.landing)),alpha =
0.5)

ggarrange(g_height,g_ground, g_craft,ncol = 1, nrow = 3)

exp(BIC.model$coefficients)[-1]
```

- ***Conclusion***
  - Distribution of Response Variable – Long Landing

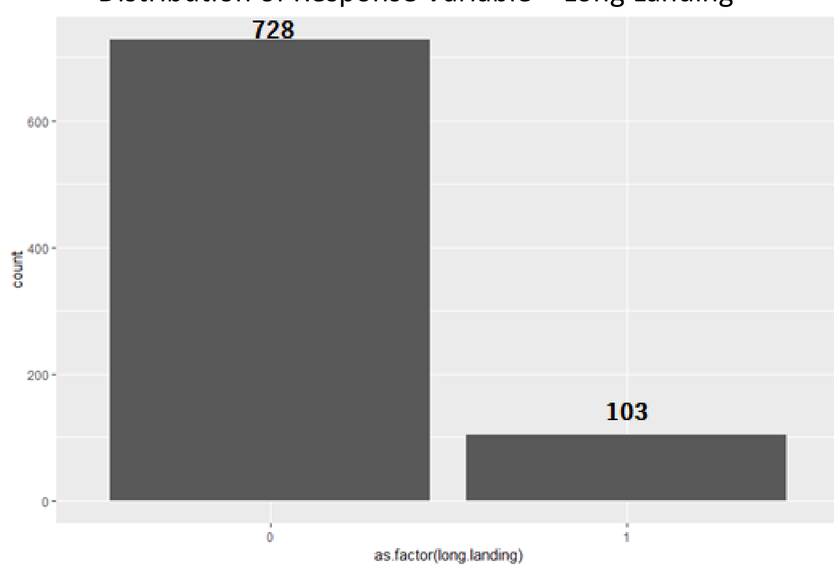

Figure 6 Distribution of Response Variable – Long Landing

  - Model - The Long Landing is significantly impacted by Speed Ground, Aircraft and Height. Other predictors are not statistically significant

- Figures – Distribution of Long Landing against significant predictors

Figure 7 Distribution of Long Landing against significant predictors

- Odds Ratio of Predictors

Predictor	Odds Ratio	Rank
aircraftboeing	155.731695	1
speed_ground	2.525837	2
height	1.259933	3

Table 5 Odds Ratio for Predictors

- Conclusion
- When we switch the make of Aircraft from Airbus to Boeing, the odds in favor of long landing increase drastically. As also visualized, Boeing has more Long landing
- When speed ground is increased by 1 mph, the chances of long landing increase by 150%. As also seen in graph, when Speed Ground increases beyond 90 mph, the landing is long
- When Height is increased by 1 mph, the chances of long landing increase by 26%. As also observed in graph, the relation between Height and Long landing is not as strong.

Step 9. Important factors for Risky Landing

Repeat Steps 1-7 but using “risky.landing” as the binary response.

- ***R code***

```
######## Risky landing

#step 2 hist

library(ggplot2)
ggplot(data = faa, aes(x = as.factor(risky.landing))) +
  geom_histogram(stat =    "count")

table(faa$risky.landing)

#step 3
t(names(faa))
faa$aircraft <- as.factor(faa$aircraft)
var_name <- rep('',7)
coeff <- rep(0,7)
odds_ratio <- rep(0,7)
direction <- rep('+',7)
p_val <- rep(0,7)
j <- 1

for(i in c(1,2,3,4,5,6,7)) {
  fit <- glm(risky.landing ~ faa[,i],family=binomial(link='logit'),data=faa)
  var_name[j] <- names(faa)[i]
  coeff[j] <- abs(summary(fit)$coefficients[2,1])
  odds_ratio[j] <- exp(fit$coefficients[2])
  if(summary(fit)$coefficients[2,1] < 0) {direction[j] <- '-'}
  p_val[j] <- summary(fit)$coefficients[2,4]
```

```r
  j <- j+1
}
tt <- cbind(1:7,var_name,coeff,odds_ratio,direction,p_val)

#step 4
library(car)
names(faa)
scatterplotMatrix(~risky.landing + no_pasg + speed_ground +
                    speed_air + height + pitch + duration, data <- faa,
                  regLine = F, ellipse = F, diagonal = F,smooth = F   )

library(ggpubr)
g_ground <- ggplot(data <- faa,aes(x=speed_ground,fill=factor(risky.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                                               colour=factor(risky.landing)),alpha =
0.5)

g_air <- ggplot(data <- faa,aes(x=speed_air,fill=factor(risky.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                                               colour=factor(risky.landing)),alpha =
0.5)

g_craft <- ggplot(data <- faa,aes(x=factor(aircraft),y=factor(risky.landing)))+
  geom_jitter(position="jitter",aes(colour=factor(risky.landing)),alpha = 0.5);

ggarrange(g_air,g_ground, g_craft,ncol = 3, nrow = 1)

#step 5
fit <- glm(risky.landing ~ aircraft + speed_ground ,
           family=binomial(link='logit'),data=faa)

summary(fit)

#step 6
null.model <- glm(risky.landing ~ 1 , family=binomial(link='logit'),data=faa)
full.model <- glm(risky.landing ~ aircraft + no_pasg + speed_ground + height +
                    pitch+ duration ,family=binomial(link='logit'),data=faa)
AIC.model.r <- step(null.model, scope=list(lower=null.model, upper=full.model),
                  direction='forward',k=2)
summary(AIC.model.r)

BIC(AIC.model.r)

#step 7
BIC.model.r <- step(null.model, scope=list(lower=null.model, upper=full.model),
                  direction='forward',k=log(nrow(faa)))
summary(BIC.model.r)
BIC(BIC.model.r)
```

- ***Code Output***

```
#step 5 model for risky landing based on our understanding
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.24398  -0.00011   0.00000   0.00000   1.61021

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -102.0772    24.7751  -4.120 3.79e-05 ***
aircraftboeing    4.0190     1.2494   3.217   0.0013 **
speed_ground      0.9263     0.2248   4.121 3.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 436.043  on 830  degrees of freedom
Residual deviance:  40.097  on 828  degrees of freedom
AIC: 46.097

Number of Fisher Scoring iterations: 12
```

- ***Observations***

We study the distribution of just risky landing to identify which is more likely in generally

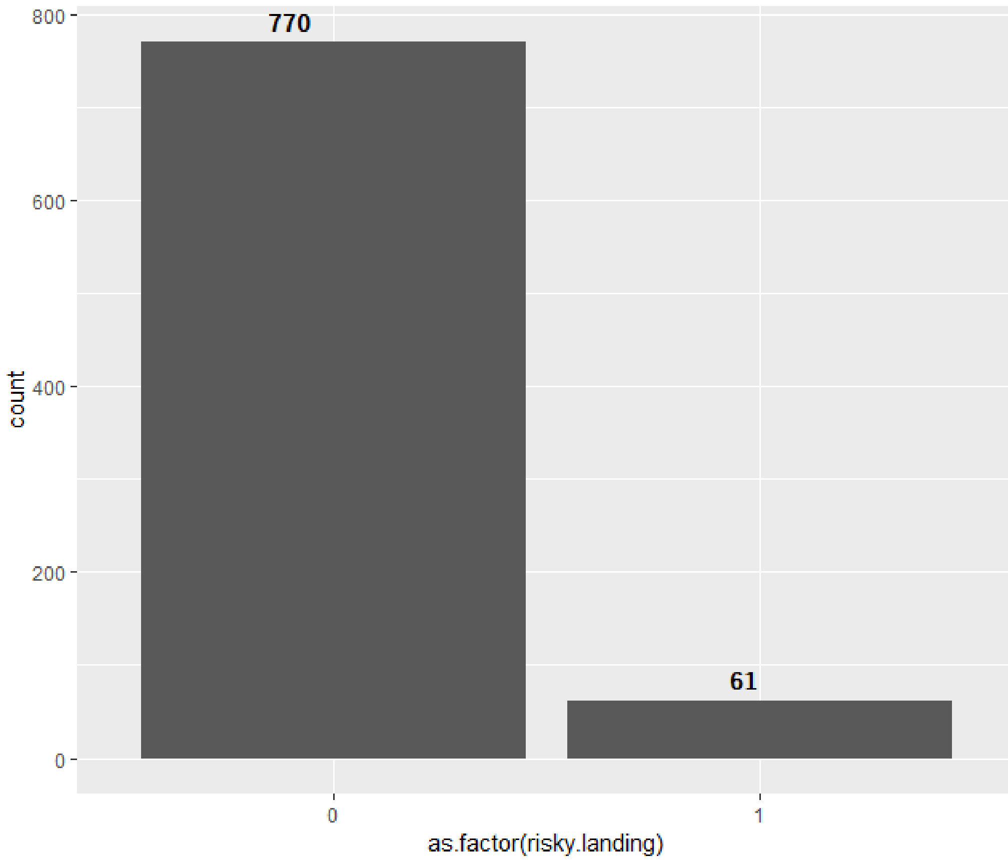

Table 6 Distribution of Risky Landing

Next, we study the individual impact of each predictor on Risky landing. It seems only Speed Ground, Speed Air and Aircraft are significant.

Sr	Variable	Coeff	Odds Ratio	Direction	P value	Rank
3	speed_ground	0.614218747	1.848212114	+	6.90E-08	1
4	speed_air	0.870401902	2.38787035	+	3.73E-06	2
1	aircraft	1.00177533	2.723111962	+	4.56E-04	3
6	pitch	0.371071969	1.449287373	+	0.143296135	4
2	no_pasg	0.025379344	0.974940004	-	0.153623692	5
7	duration	0.001151836	0.998848827	-	0.680198706	6
5	height	0.002218606	0.997783854	-	0.870591704	7

Table 7 P value ranks for Risky landing

Now we study the distribution of Risky Landing against significant variables. There is a clear distinction for Speed Ground and Speed Air as to when a landing may become risky.

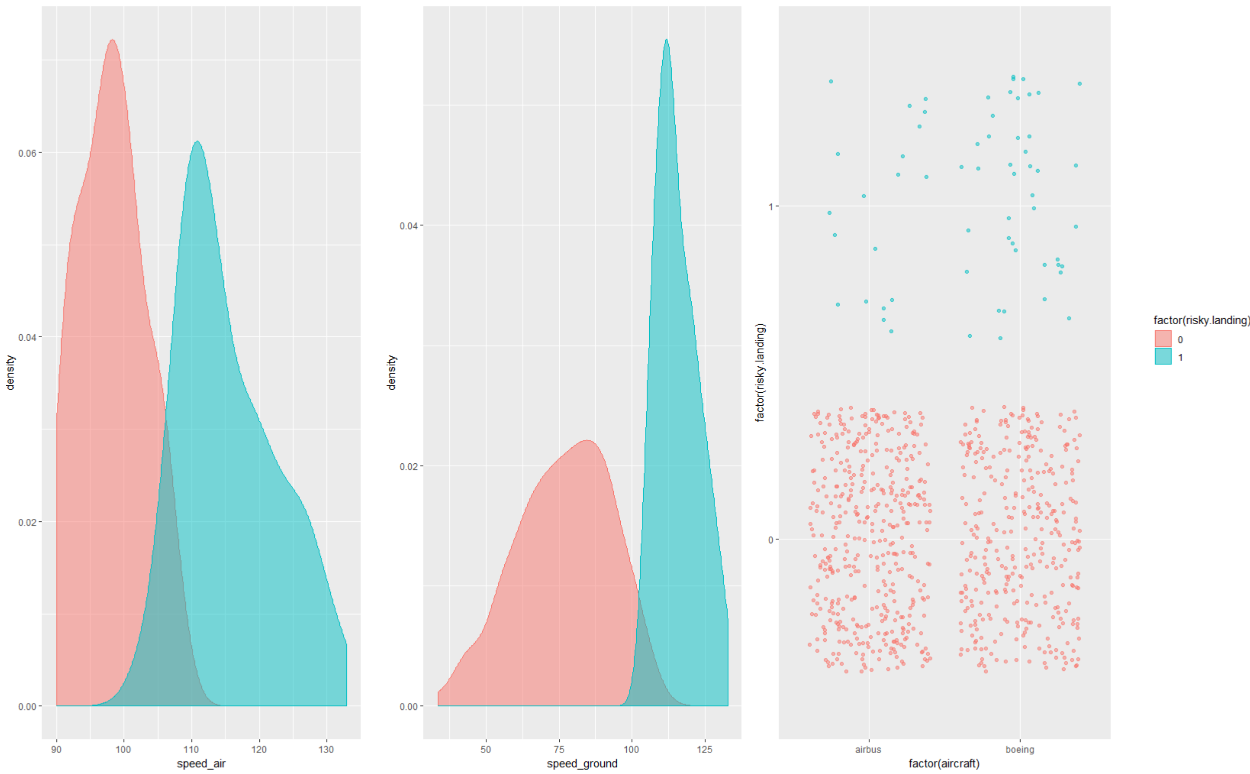

Figure 8 Distribution of Risky Landing by significant predictors

As we are already aware, Speed Air has missing values and is highly correlated with Speed Ground, we will only be using the 2 significant variables we arrived at i.e. Speed Ground and Aircraft. Now we study the comparison of AIC Forward vs BIC Forward model selection for Risky Landing. Here, we notice that the model decided based on individual significance i.e. Speed Ground & Aircraft for prediction, is like the model obtained through BIC forward step selection and hence this will be our final model for Risky Landing.

Properties	**Step Forward AIC for Risky Landing**					**Step Forward BIC for Risky Landing**				
Deviance Residuals	**Min**	**1Q**	**Median**	**3Q**	**Max**	**Min**	**1Q**	**Median**	**3Q**	**Max**
	-2.33913	-0.00009	0	0	1.8781	-2.24398	-0.00011	0	0	1.61021
Coefficients	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**		**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	-99.9078	25.57993	-3.906	**0.0000939**		-102.0772	24.7751	-4.12	**0.0000379**	
speed_ground	0.94963	0.23559	4.031	**0.0000556**		0.9263	0.2248	4.121	**0.0000378**	
aircraftboeing	4.64188	1.4752	3.147	**0.00165**		4.019	1.2494	3.217	**0.0013**	
no_pasg	-0.08462	0.05732	-1.476	**0.13987**		NA				
Null deviance on degrees of freedom	436.043 on 830					436.043 on 830				
Residual deviance on degrees of freedom	37.707 on 827					40.097 on 828				
AIC	45.707					46.097				
BIC	64.59746					60.26449				
Number of Fisher Scoring iterations	12					12				

Table 8 AIC vs BIC Model selection criteria for Risky landing

- ***Conclusion***

Based on the iterative approach discussed under the observation section, the final model selected is Risky Landing regressed on Speed Ground and Aircraft. Below are the statistics for the final model.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.24398  -0.00011   0.00000   0.00000   1.61021

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -102.0772    24.7751  -4.120 3.79e-05 ***
speed_ground      0.9263     0.2248   4.121 3.78e-05 ***
aircraftboeing    4.0190     1.2494   3.217   0.0013 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 436.043  on 830  degrees of freedom
Residual deviance:  40.097  on 828  degrees of freedom
AIC: 46.097

Number of Fisher Scoring iterations: 12
```

Step 10. Intermediate Summary of findings for Risky Distance

You are scheduled to meet with an FAA agent who wants to know "what are risk factors for risky landings and how do they influence its occurrence?". For your presentation, you are only allowed to show. The question is: what model/table/figures/statements you would include in your presentation. Be selective! One model, One table, No more than three figures & No more than five bullet statements. Please use statements that she can understand.

- Distribution of Response Variable – Risky Landing

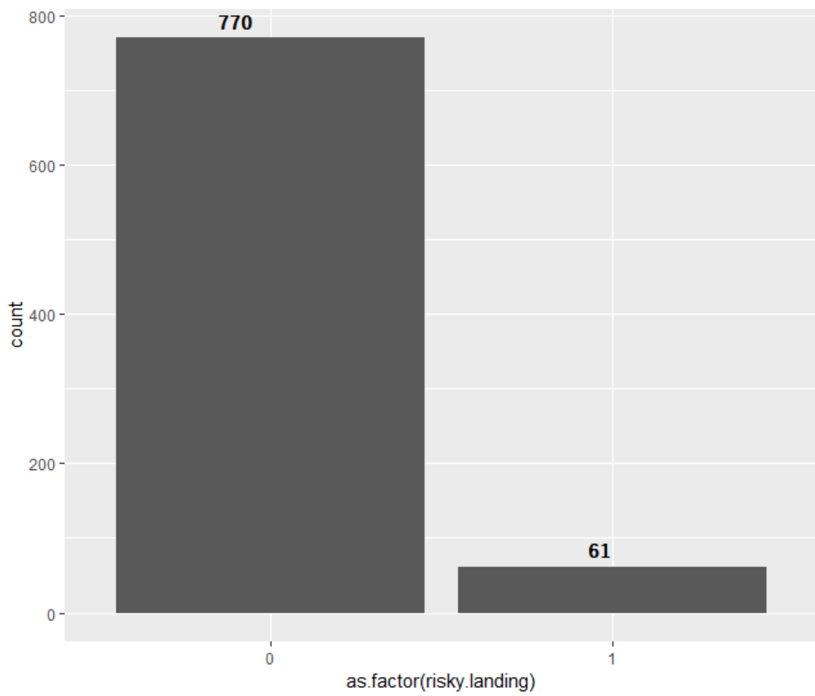

Figure 9 Distribution of Risky Landing

- **Model** - The Risky Landing is significantly impacted by Speed Ground & Aircraft. Other predictors are not statistically significant
- **Figures** – Distribution of Risky Landing against significant predictors

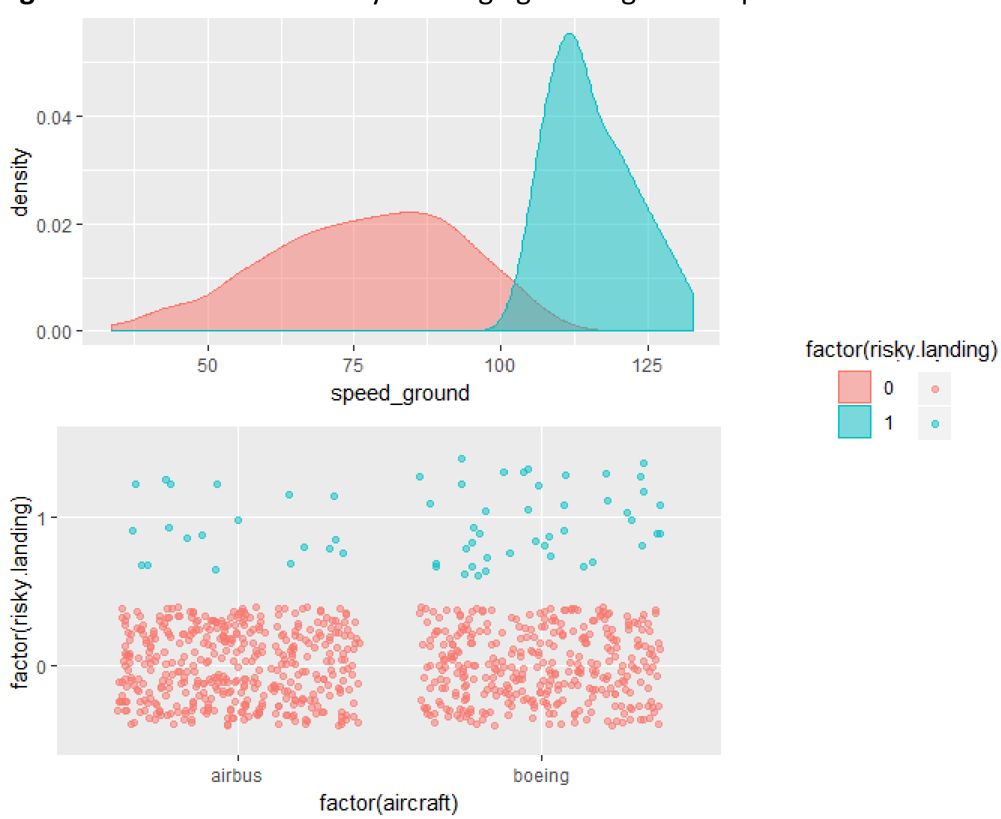

Figure 10 Distribution of Risky Landing by significant factors

- **Odds Ratio of Predictors**

Predictor	Odds Ratio	Rank
aircraftboeing	**55.647166**	**1**
speed_ground	**2.525084**	**2**

Table 9 Odds Ratio for Risky Landing

- **Conclusion**
  1. When we switch the make of Aircraft from Airbus to Boeing, the odds in favor of Risky landing increase drastically. As also visualized, Boeing has more Long landing
  2. When speed ground is increased by 1 mph, the chances of Risky landing increase by 152%. As also seen in graph, when Speed Ground increases beyond 100 mph, the landing is risky

Step 11. Model Comparison for Long Landing & Risky Landing

Use no more than three bullet statements to summarize the difference between the two models.

- ***Observations***

<table>
<tr><th>Properties</th><th colspan="5">Step Forward BIC for Long Landing</th><th colspan="5">Step Forward BIC for Risky Landing</th></tr>
<tr><td rowspan="2">Deviance Residuals</td><td>Min</td><td>1Q</td><td>Median</td><td>3Q</td><td>Max</td><td>Min</td><td>1Q</td><td>Median</td><td>3Q</td><td>Max</td></tr>
<tr><td>-2.43442</td><td>-
0.00117</td><td>0</td><td>0</td><td>2.5744</td><td>-2.24398</td><td>-0.00011</td><td>0</td><td>0</td><td>1.61021</td></tr>
<tr><td>Coefficients</td><td>Estimate</td><td>Std. Error</td><td>z value</td><td colspan="2">Pr(>|z|)</td><td>Estimate</td><td>Std. Error</td><td>z value</td><td colspan="2">Pr(>|z|)</td></tr>
<tr><td>(Intercept)</td><td>-103</td><td>19.229</td><td>-5.35</td><td colspan="2">8.59E-08</td><td>-102.0772</td><td>24.7751</td><td>-4.12</td><td colspan="2">0.0000379</td></tr>
<tr><td>speed_ground</td><td>0.9266</td><td>0.1724</td><td>5.374</td><td colspan="2">7.7E-08</td><td>0.9263</td><td>0.2248</td><td>4.121</td><td colspan="2">0.0000378</td></tr>
<tr><td>aircraftboeing</td><td>5.0481</td><td>1.1152</td><td>4.527</td><td colspan="2">6E-06</td><td>4.019</td><td>1.2494</td><td>3.217</td><td colspan="2">0.0013</td></tr>
<tr><td>height</td><td>0.2311</td><td>0.0596</td><td>3.877</td><td colspan="2">0.00011</td><td colspan="5">NA</td></tr>
<tr><td>Null deviance on degrees of freedom</td><td colspan="5">622.778 on 830</td><td colspan="5">436.043 on 830</td></tr>
<tr><td>Residual deviance on degrees of freedom</td><td colspan="5">57.047 on 827</td><td colspan="5">40.097 on 828</td></tr>
<tr><td>AIC</td><td colspan="5">65.047</td><td colspan="5">46.097</td></tr>
<tr><td>BIC</td><td colspan="5">83.9372</td><td colspan="5">60.26449</td></tr>
<tr><td>Number of Fisher Scoring iterations</td><td colspan="5">11</td><td colspan="5">12</td></tr>
</table>

Table 10 Step Forward BIC models for Long vs Risky Landing

<table>
<tr><th rowspan="2">Predictor</th><th>Long Landing</th><th></th><th>Risky Landing</th><th></th></tr>
<tr><th>Odds Ratio</th><th>Rank</th><th>Odds Ratio</th><th>Rank</th></tr>
<tr><td>aircraftboeing</td><td>155.731695</td><td>1</td><td>55.6472</td><td>1</td></tr>
<tr><td>speed_ground</td><td>2.525837</td><td>2</td><td>2.52508</td><td>2</td></tr>
<tr><td>height</td><td>1.259933</td><td>3</td><td colspan="2">NA</td></tr>
</table>

Table 11 Odds Ratio for significant Predictors

- ***Conclusion***
  1. Both models are built on BIC as Forward selection criteria
  2. Height is a predictor for Long Landing but not Risky Landing
  3. Speed Ground Impacts both the Responses in a similar way
  4. When the Aircraft is Boeing, the chances of a Long landing are increased by a greater percentage as they would for Risky landing

Step 12. ROC Curve

Plot the ROC curve (sensitivity versus 1-specificity) for each model (see pp.32-33 in Lecture 4 slides). Draw the two curves in the same plot. Do you have any comment?

- ***R code***

```
#step 12
pred.l <- ifelse(predict(BIC.model,type = 'response') < 0.5,0,1)
pred.r <- ifelse(predict(BIC.model.r,type = 'response') < 0.5,0,1)

thresh <- seq(0.01,0.5,0.01)
sensitivity <- specificity <- rep(NA,length(thresh))
for( j in seq(along=thresh)) {
  pp<- ifelse(predict(BIC.model,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~faa$long.landing+pp)
  specificity[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}
par(mfrow=c(1,2))
matplot(thresh,cbind(sensitivity,specificity),type="l",xlab="Threshold",
        ylab="Proportion",lty=1:2)
plot(1-specificity,sensitivity,type="l");abline(0,1,lty=2)


pred.r <- ifelse(predict(BIC.model.r,type = 'response') < 0.5,0,1)

thresh <- seq(0.01,0.5,0.01)
sensitivity.r <- specificity.r <- rep(NA,length(thresh))
for( j in seq(along=thresh)) {
  pp<- ifelse(predict(BIC.model.r,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~faa$risky.landing+pp)
  specificity.r[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity.r[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}
par(mfrow=c(1,2))
matplot(thresh,cbind(sensitivity.r,specificity.r),type="l",xlab="Threshold",
        ylab="Proportion",lty=1:2)
plot(1-specificity.r,sensitivity.r,type="l");abline(0,1,lty=2)

plot(1-specificity,sensitivity, type="l", col="blue")
points(1-specificity.r,sensitivity.r,type="l",col="red")
lines(1-specificity.r,sensitivity.r, col="red",lty=2)
```

- ***Code Output***

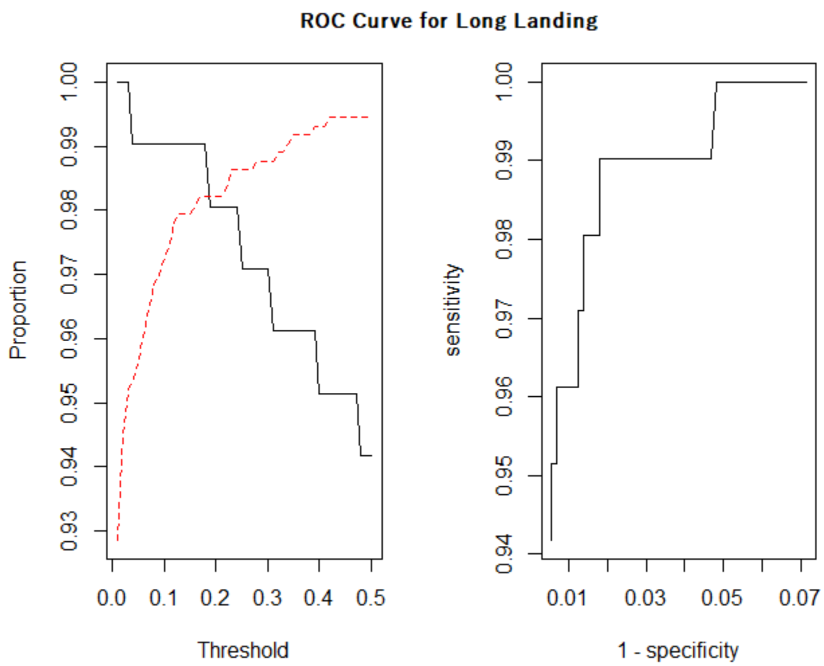

Figure 11 ROC Curve for Long Landing

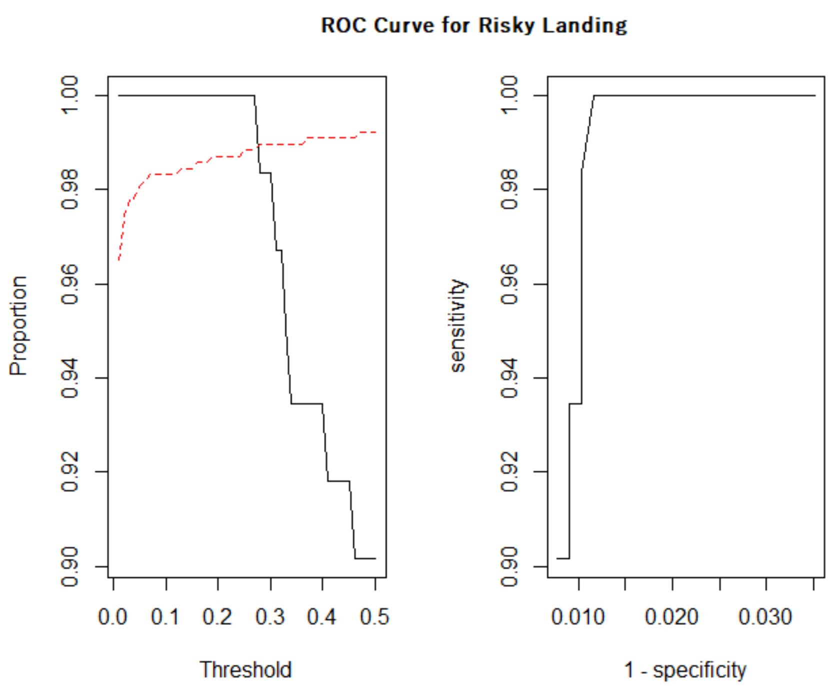

Figure 12 Roc Curve for Risky Landing

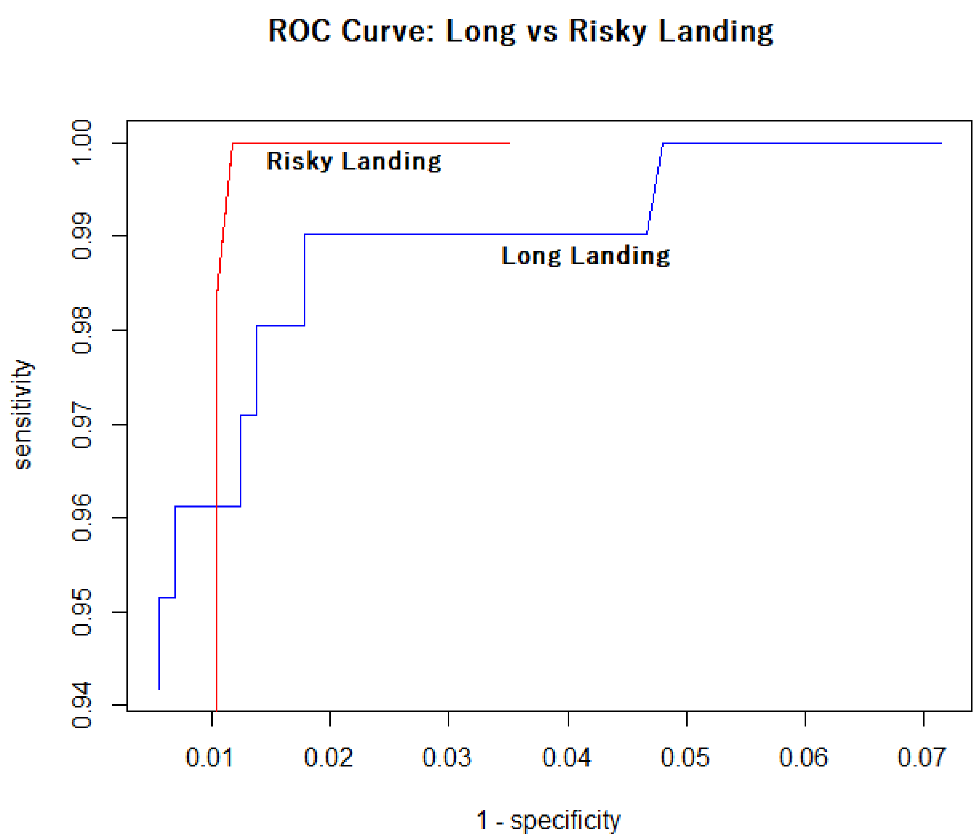

Figure 13 ROC Curve for Long vs Risky Landing

- ***Observations***

We observe the variation of proportion of Long landing and Risky landing versus a given threshold probability. We also study the ROC curve for the 2 responses based on their models.

- ***Conclusion***

ROC Curve shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. In our analysis, when we plot both the curves in the same plot, we can do a comparative study of the prediction power of both models for their respective responses. We see that Risky Landing is predicted with more accuracy as compared to Long Landing.

Step 13. Model Prediction

A commercial airplane is passing over the threshold of the runway, at this moment we have its basic information and measures of its airborne performance (Boeing, duration=200, no_pasg=80, speed_ground=115, speed_air=120, height=40, pitch=4). Predict its probability of being a long landing and a risky landing, respectively. Report the predicted probability as well as its 95% confidence interval.

- ***R code***

```
#step13
new.ind <- data.frame(aircraft="boeing",duration=200,no_pasg=80,
                      speed_ground=115,speed_air=120,height=40,pitch=4)
p.l <- predict(BIC.model,newdata=new.ind,type = 'response',se.fit = T)
p.r <- predict(BIC.model.r,newdata=new.ind,type='response' ,se.fit = T)

c(p.l$fit,p.l$fit-2*p.l$se.fit[1],p.l$fit+2*p.l$se.fit[1])

c(p.r$fit,p.r$fit-2*p.r$se.fit[1],p.r$fit+2*p.r$se.fit[1])
```

- ***Observations***

Response	Probability	Lower Limit	Upper Limit
Long Landing	1	0.9999999	1.0000001
Risky Landing	0.999789	0.9989074	1.0006706

Table 12 Landing Prediction using BIC model

- ***Conclusion***

We use the predict() function to make the prediction for desired responses. We notice that the probabilities for Long Landing and Risky Landing are very close to 1 with a very low standard error which further narrows down the confidence interval. Regardless of the cutoff threshold, I think both responses are set to 1. This result is also logically coherent because if the flight is predicted as Risky, it is already Long.

Step 14. Compare models with different link functions

For the binary response "risky landing", fit the following models using the risk factors identified in Steps 9-10:

- Probit model
- Hazard model with complementary log-log link

Compare these two models with the logistic model. Do you have any comments?

- ***R code***

```
#step14
r.logit <- glm(risky.landing ~ speed_ground + aircraft,
                  data = faa, family=binomial(link='logit'))
summary(r.logit)

r.probit <- glm(risky.landing ~ speed_ground + aircraft,
                  data = faa, family=binomial(link='probit'))
summary(r.probit)
```

```
r.haz <- glm(risky.landing ~ speed_ground + aircraft,
                      data = faa, family=binomial(link='cloglog'))
summary(r.haz)
```

- ***Code Output***

Properties	Logit					Probit					Hazard (Cloglog)				
Deviance Residuals	Min	1Q	Median	3Q	Max	Min	1Q	Median	3Q	Max	Min	1Q	Median	3Q	Max
	-2.24398	-0.00011	0	0	1.61021	-2.21	0	0	0	1.573	-2.24103	-0.00183	-0.00004	0	1.67963
Coefficients	Estimate	Std. Error	z value	Pr(>|z|)		Estimate	Std. Error	z value	Pr(>|z|)		Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-102.0772	24.7751	-4.12	0.0000379		-58.6931	13.3133	-4.409	1.04E-05		-69.2654	14.7396	-4.699	0.00000261	
speed_ground	0.9263	0.2248	4.121	0.0000378		0.5322	0.1207	4.411	1.03E-05		0.6221	0.1326	4.69	0.00000274	
aircraftboeing	4.019	1.2494	3.217	0.0013		2.3567	0.7016	3.359	0.000782		2.8984	0.8002	3.622	0.000292	
Null deviance on df	436.043 on 830					436.043 on 830					436.043 on 830				
Residual deviance on df	40.097 on 828					39.436 on 828					41.443 on 828				
AIC	46.097					45.436					47.443				
BIC	60.26449					59.60437					61.6113				
Fisher Scoring iterations	12					14					13				

Table 13 Logit, Probit & Hazard comparison for Risky Landing

- ***Conclusion***

1. Deviance Residual for Probit Model is the best
2. AIC, BIC criteria if, it was to be applied, it would have suggested Probit model
3. The Standard Error for Probit Model is the least
4. Logit model has the maximum size of coefficients
5. Probit Model and Clog log model are closer in terms of their estimates as compared to Clog log model
6. To conclude, we will prefer Probit model over Logit model based on the summary of model

Step 15. ROC Curve comparison

Compare the three models by showing their ROC curves in the same plot (see Step 12).

- ***R code***

```
#step 15
plot(1-specificity.l,sensitivity.l, type="l", col="blue")

#points(1-specificity.p,sensitivity.p,type="o",col="red",pch=21)
lines(1-specificity.p,sensitivity.p, type = "b",col="red",lty=4)

#points(1-specificity.c,sensitivity.c,type="x",col="green",pch=25)
lines(1-specificity.c,sensitivity.c, type = "o",col="green",lty=3)

par(mfrow=c(1,3))
plot(1-specificity.l,sensitivity.l, type="l", col="blue",main = 'Logit')
plot(1-specificity.p,sensitivity.p, type="l", col="red",main = 'Probit')
plot(1-specificity.c,sensitivity.c, type="l", col="green",main = 'Hazard')
```

- ***Code Output***

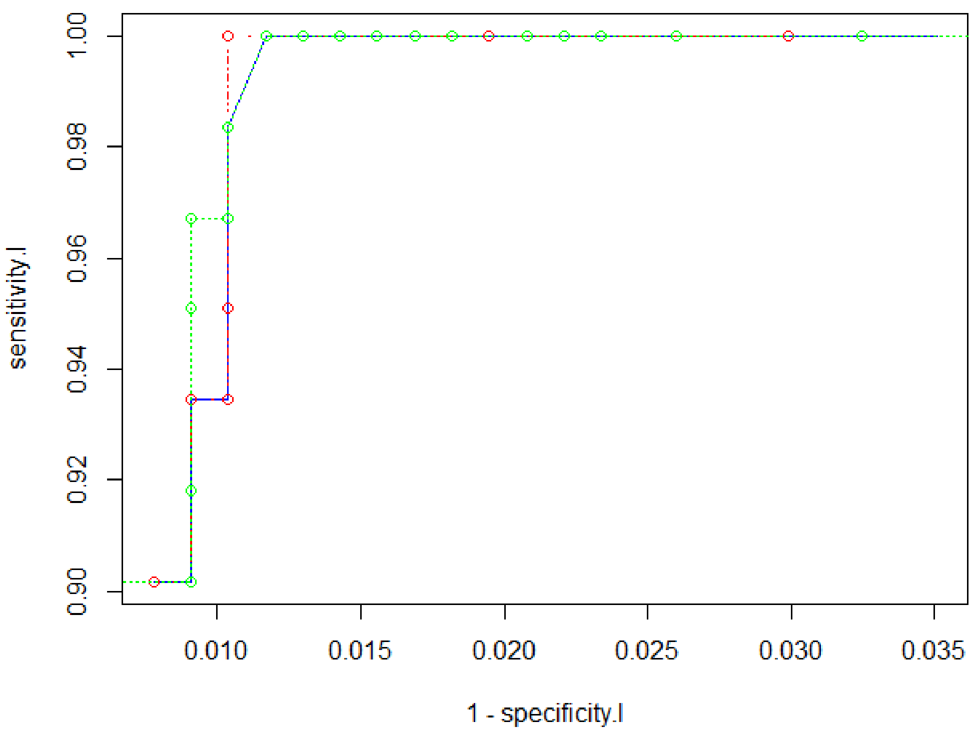

Table 14 Combined ROC Curve for Risky Landing

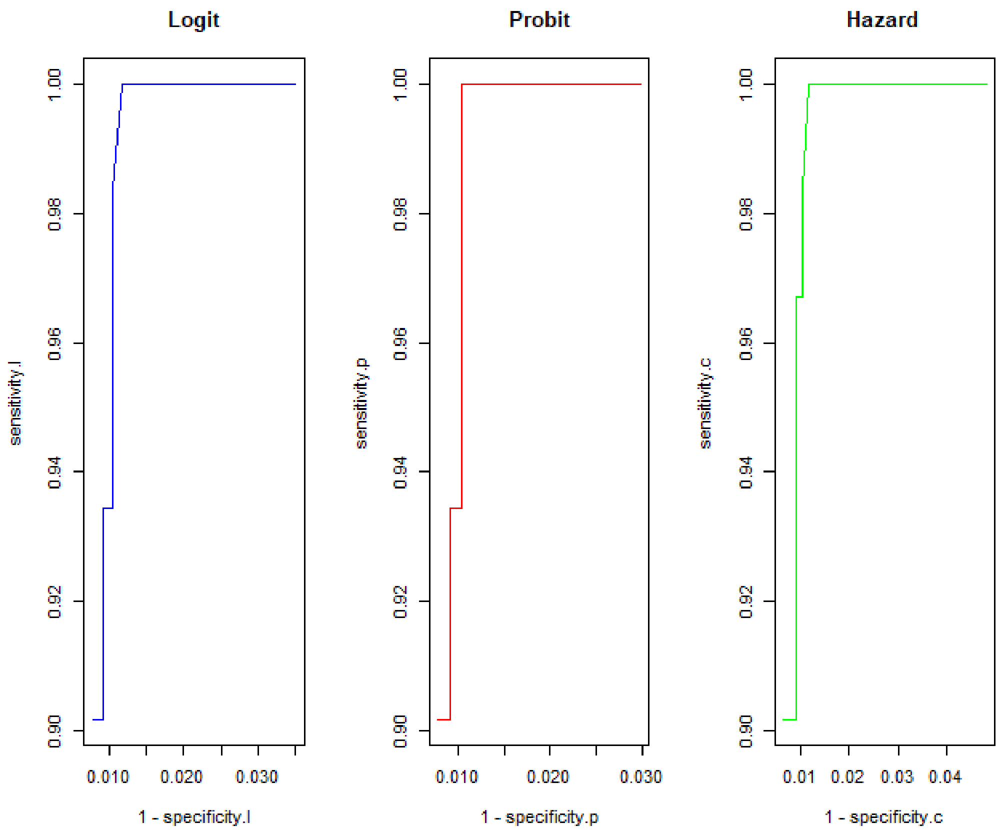

Table 15 ROC Curve of different models for RIsky Landing

- ***Observations***

I tried plotting all 3 ROC Curves on the same graph in order to better visualize their differences. However, there was such overlap that I was not able to decipher it. Hence, I went with plotting 3 separate curves aligned by sensitivity values.

- ***Conclusion***

Although there is much overlap between the 3 curves, we can see that Hazard model has the maximum area under the curve which indicates it is most suitable for high prediction accuracy. Also, the AUC for Logit and Probit is very similar which indicates there is not much difference in their predictive powers.

Step 16. Top-N Outcomes' Comparison

Use each model to identify the top 5 risky landings. Do they point to the same flights?

- ***R code***

```
#step 16
```

```
pred.logit <- predict(r.logit,type = 'response')
pred.probit <- predict(r.probit,type = 'response')
pred.hazard <- predict(r.haz,type = 'response')

pred.logit[which(pred.logit==max(pred.logit))]


faa[as.numeric(names(tail(sort(pred.logit),5))),] # Logit Model
faa[as.numeric(names(tail(sort(pred.probit),5))),] # Probit Model
faa[as.numeric(names(tail(sort(pred.hazard),5))),] # Hazard Model

top.logit <- sort(as.numeric(names(tail(sort(pred.logit),5))))
top.probit <- sort(as.numeric(names(tail(sort(pred.probit),5))))
top.hazard <- sort(as.numeric(names(tail(sort(pred.hazard),5))))

print(topn <- cbind(top.logit,top.probit,top.hazard))
```

- ***Code Output***

```
> faa[as.numeric(names(tail(sort(pred.logit),5))),] # Logit Model
    aircraft no_pasg speed_ground speed_air   height    pitch  duration long.landing
risky.landing
227   airbus      60     131.0352  131.3379 28.27797 3.660194 131.73110            1
1
675   boeing      61     126.8393  126.1186 20.54783 4.334558 153.83445            1
1
814   boeing      72     129.2649  128.4177 33.94900 4.139951 161.89247            1
1
773   boeing      67     129.3072  127.5933 23.97850 5.154699 154.52460            1
1
513   boeing      52     132.7847  132.9115 18.17703 4.110664  63.32952            1
1
> faa[as.numeric(names(tail(sort(pred.probit),5))),] # Probit Model
    aircraft no_pasg speed_ground speed_air   height    pitch duration long.landing
risky.landing
675   boeing      61     126.8393  126.1186 20.54783 4.334558 153.8345            1
1
772   boeing      67     122.7566  123.8826 30.21657 3.213703 116.9845            1
1
773   boeing      67     129.3072  127.5933 23.97850 5.154699 154.5246            1
1
784   boeing      68     126.6692  127.9641 23.76423 2.993151 197.5464            1
1
814   boeing      72     129.2649  128.4177 33.94900 4.139951 161.8925            1
1
> faa[as.numeric(names(tail(sort(pred.hazard),5))),] # Hazard Model
    aircraft no_pasg speed_ground speed_air   height    pitch duration long.landing
risky.landing
760   boeing      66     117.6406  112.2650 35.91004 4.058218 109.4517            1
1
772   boeing      67     122.7566  123.8826 30.21657 3.213703 116.9845            1
1
773   boeing      67     129.3072  127.5933 23.97850 5.154699 154.5246            1
1
```

```
784    boeing       68      126.6692  127.9641 23.76423 2.993151 197.5464            1
1
814    boeing       72      129.2649  128.4177 33.94900 4.139951 161.8925            1
1
```

- ***Observations***
  - Highlight rules for above Index table are as follows:
  - Red – Common in none
  - Blue – Common in Logit and Probit
  - Yellow – Common in Probit and Hazard
  - Green – Common in all 3 models

Rank	top.logit	top.probit	top.hazard
1	227	675	760
2	513	772	772
3	675	773	773
4	773	784	784
5	814	814	814

Table 16 Indexes of Top 5 predictions from different models

- ***Conclusion***

We notice that there are 2 similar predictions for all the 3 models. However, if I checked the prediction probability for them and it was 1. Basically, there a lot more ones in all the 3 models' predicted probabilities which might induce some randomness as to which are the top 5 when chosen for displaying. On the other hand, all the top-5 predictions from all models have high-speed Ground values and are Type = Boeing which intuitively suggest us that it is very likely that the landing is risky for them

Step 17. Confidence interval for different models

Use the Probit model and hazard model to make prediction for the flight described in Step 13. Report the predicted probability as well as its 95% confidence interval. Compare the results with that from Step 13.

- ***R code***

```
# step 17
r.logit <- glm(risky.landing ~ speed_ground + aircraft,
               data = faa, family=binomial(link='logit'))
r.probit <- glm(risky.landing ~ speed_ground + aircraft,
                data = faa, family=binomial(link='probit'))
r.haz <- glm(risky.landing ~ speed_ground + aircraft,
             data = faa, family=binomial(link='cloglog'))
l.logit <- glm(long.landing ~ speed_ground + aircraft + height,
               data = faa, family=binomial(link='logit'))
```

```r
l.probit <- glm(long.landing ~ speed_ground + aircraft + height,
                data = faa, family=binomial(link='probit'))
l.haz <- glm(long.landing ~ speed_ground + aircraft + height,
             data = faa, family=binomial(link='cloglog'))


new.ind <- data.frame(aircraft="boeing",duration=200,no_pasg=80,
                      speed_ground=115,speed_air=120,height=40,pitch=4)
r.l.predict <- predict(r.logit,newdata=new.ind,type = 'response',se.fit = T)
r.p.predict <- predict(r.probit,newdata=new.ind,type='response' ,se.fit = T)
r.h.predict <- predict(r.haz,newdata=new.ind,type='response' ,se.fit = T)
l.l.predict <- predict(l.logit,newdata=new.ind,type = 'response',se.fit = T)
l.p.predict <- predict(l.probit,newdata=new.ind,type='response' ,se.fit = T)
l.h.predict <- predict(l.haz,newdata=new.ind,type='response' ,se.fit = T)

p_vector <- c(r.l.predict$fit,r.p.predict$fit,r.h.predict$fit,l.l.predict$fit,
              l.p.predict$fit,l.h.predict$fit)
se_vector <- c(r.l.predict$se.fit,r.p.predict$se.fit,r.h.predict$se.fit,
               l.l.predict$se.fit,l.p.predict$se.fit,l.h.predict$se.fit)
n_vector <- c("Risky logit", "Risky probit","Risky hazard", "Long logit",
              "Long probit","Long hazard")

tt <- cbind(p_vector,se_vector,n_vector)
```

- ***Code Output***

Statistic	Risky			Long		
	Logit	Probit	Hazard	Logit	Probit	Hazard
Probability	0.999788977	0.999999448	1	0.999999983	1	1
Std. Err	0.000440811	3.15356E-06	2.60552E-16	5.87047E-08	3.86266E-16	4.31668E-16
Confidence Lower	0.998907354	0.999993141	1	0.999999866	1	1
Confidence Upper	1.0006706	1.000005755	1	1.0000001	1	1

Table 17 Prediction comparison for different types of models for Risky and Long Landing

- ***Observations***

The interval for Hazard and Probit Predictions are very narrow, as narrow as that they are not observable.

- ***Conclusion***

All models make similar prediction that the flight is going to have a long and risky landing. Hence, we conclude that there isn't much difference in the prediction power of the models for this data point. However, this does not establish their similarities in other scenarios