Project1 Part2

Tess Newkold

2/17/2019

Part 1 - code and output not shown

Create Binary Responses

Step 1

```
#creating binary variables for longLanding and riskyLanding
cleanFAA$longLanding <- ifelse(cleanFAA$distance > 2500, 1, 0)
cleanFAA$riskyLanding <- ifelse(cleanFAA$distance > 3000, 1, 0)

#get rid of the distance column
FAA <- subset(cleanFAA, select = -c(distance))
```

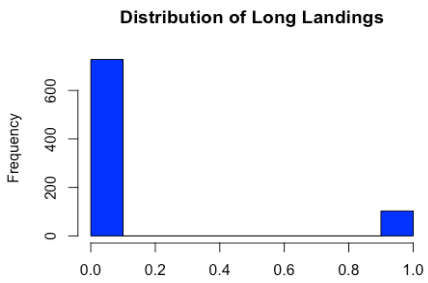
Conclusion: Two new columns were created, 'longLanding' for landing distances longer than 2500 feet, and 'riskyLanding' for landings longer than 3000 feet. The continous column 'distance' was then discarded. We will now only use the new binary columns for analysis.

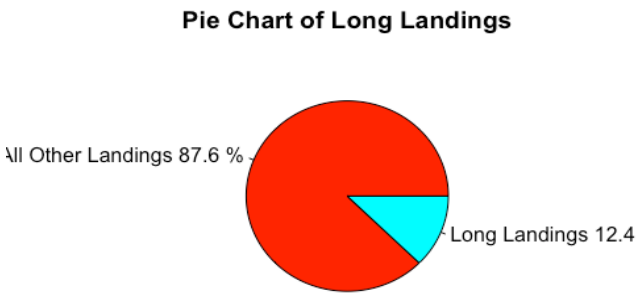
Identify Importnat Factors Using the Binary Data of longLanding

Step 2

```
#histogram showing distribution of long landings
hist(FAA$longLanding, main = "Distribution of Long Landings", col = 'blue', xlab = ' ')

pct <- round(table(FAA$longLanding) / length(FAA$longLanding) * 100, 1)
labs <- c("All Other Landings", "Long Landings")
labs <- paste(labs, pct)
labs <- paste(labs, "%", sep = " ")
pie(table(FAA$longLanding), labels = labs, col = rainbow(length(labs)), main = "Pie Chart
of Long Landings")
```





Conclusion: There are by far more landing that are not long. However, 12.4% of landings being long is quite high, so it is very important to evaluate. You can see this in both the histogram and the pie chart above.

Step 3

```
#model for no-pasg
modelNoPasgB <- glm(FAA$longLanding ~ FAA$no_pasg, family = binomial)
summary(modelNoPasgB)
#model for speed on ground
modelSpeedOnGroundB <- glm(FAA$longLanding ~ FAA$speed_ground, family = binomial)
summary(modelSpeedOnGroundB)
#model for speed in air
modelSpeedInAirB <- glm(FAA$longLanding ~ FAA$speed_air, family = binomial)
summary(modelSpeedInAirB)
#model for height
modelHeightB <- glm(FAA$longLanding ~ FAA$height,family = binomial)
summary(modelHeightB)
#model for pitch
modelPitchB <- glm(FAA$longLanding ~ FAA$pitch, family = binomial)
summary(modelPitchB)
#make dummy variable for aircraft Airbus = 0, Boeing = 1
dummy <- as.numeric(FAA$aircraft == "airbus", 0, 1)
#model for aircraft
modelAircraftB <- lm(FAA$longLanding ~ dummy, data = FAA)
summary(modelAircraftB)

#model for duration
modelDurationB <- lm(FAA$longLanding ~ duration, data = FAA)
summary(modelDurationB)

##
## Call:
## lm(formula = FAA$longLanding ~ duration, data = FAA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1415 -0.1316 -0.1270 -0.1218  0.8865
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1464658  0.0401670   3.646 0.000284 ***
## duration    -0.0001190  0.0002477  -0.481 0.630979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3345 on 779 degrees of freedom
##   (50 observations deleted due to missingness)
## Multiple R-squared:  0.0002963,  Adjusted R-squared:  -0.000987
## F-statistic: 0.2309 on 1 and 779 DF,  p-value: 0.631

#Sample output for all variables
```

Calculating the odds ratio

```
exp(modelNoPasgB$coefficients)
exp(modelSpeedOnGroundB$coefficients)
exp(modelSpeedInAirB$coefficients)
exp(modelHeightB$coefficients)
exp(modelPitchB$coefficients)
exp(modelDurationB$coefficients)
```

```
exp(modelAircraftB$coefficients)

## (Intercept)       dummy
##   1.1890166   0.9120636

#Sample output for all variables
```

Table that Ranks the factors from most important to the least

Table 1.

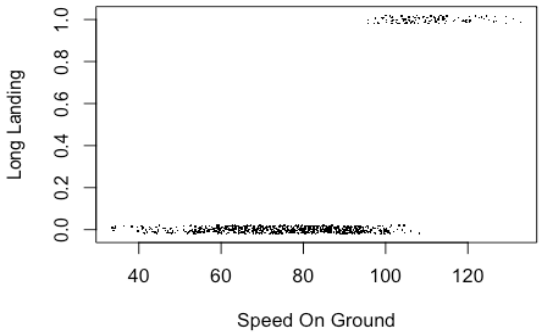
Rank	Variable	Regression Coefficient	Odd Ratio	Direction of Regression Coefficient	p-value
1	Speed On Ground	0.47235	1.603752	+	3.94e-14
2	Speed In Air	0.51232	1.669162	+	4.33e-11
3	Aircraft	-0.09205	0.9120636	-	5.57e-05
4	Pitch	0.4005	1.49261233	+	0.0466
5	Height	0.008624	1.0086613	+	0.422
6	Number of Passengers	-0.007256	0.9927699	-	0.6059
7	Duration	-0.0001190	0.999881	-	0.630979

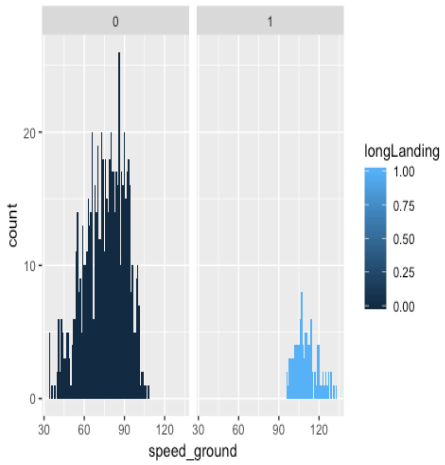
Conclusion: From the information above it appears that speed on ground is the most important factor in having a long landing. Followed by Speed In Air, Aircraft, and Pitch with the rest not looking significant.

Step 4

```
plot(jitter(longLanding,0.1)~jitter(speed_ground), FAA, xlab="Speed On Ground", ylab =
"Long Landing", pch = ".")

ggplot(FAA, aes(x = speed_ground, fill = longLanding)) + geom_histogram(position =
"dodge",binwidth = 1) +
  facet_grid(~longLanding)
```

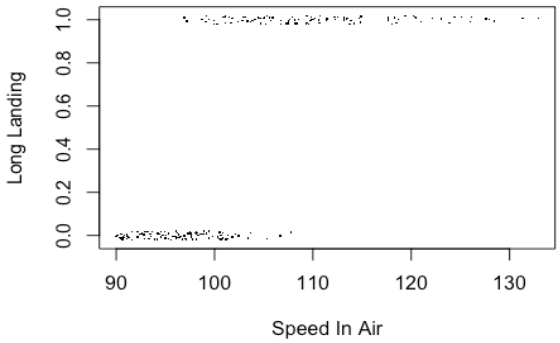


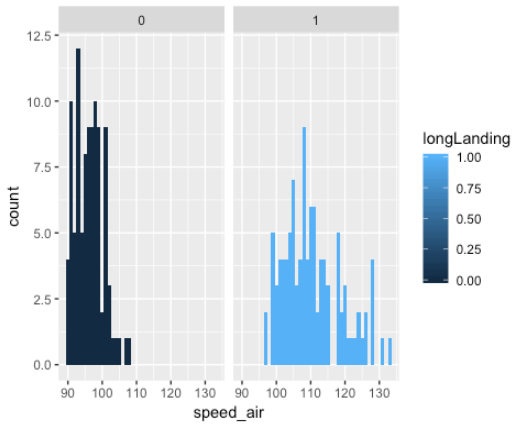


Observations: From the above graphs for 'Speed On Ground' you can see that there is a pattern between Long Landing and Speed on Ground. Most of the flights that had long landings were going at a faster speed on ground than their counterparts. Also, the distribution of speed on ground is realatively normal.

```
plot(jitter(longLanding, 0.1)~jitter(speed_air), FAA, xlab = "Speed In Air", ylab = "Long
Landing", pch = ".")

ggplot(FAA, aes(x = speed_air, fill = longLanding)) + geom_histogram(position =
"dodge",binwidth = 1) +
  facet_grid(~longLanding)
```

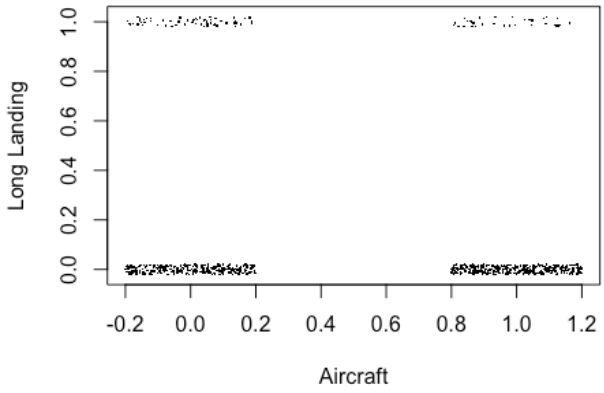


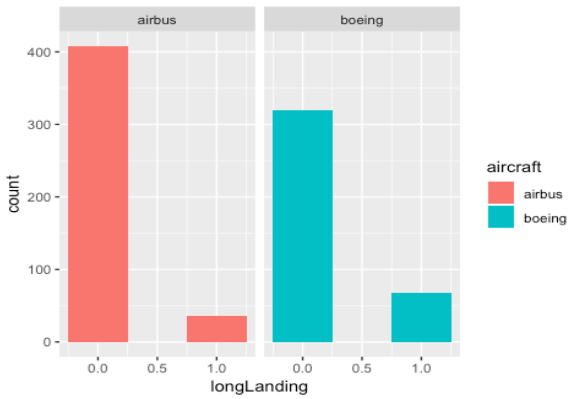


Observations: You can see that there is a pattern between Long Landing and Speed in Air. Most of the flights that had long landings were going at a faster speed on air than their counterparts, however there still are some flights that are going slower that are still going over the runway. Also, the distribution of speed on ground is realatively normal.

```
plot(jitter(longLanding, 0.1)~jitter(dummy), FAA, xlab = "Aircraft", ylab = "Long
Landing", pch = ".")

ggplot(FAA, aes(x = longLanding, fill = aircraft)) +
  geom_bar(position = "dodge", width = 0.5) +
  facet_grid(~aircraft)
```

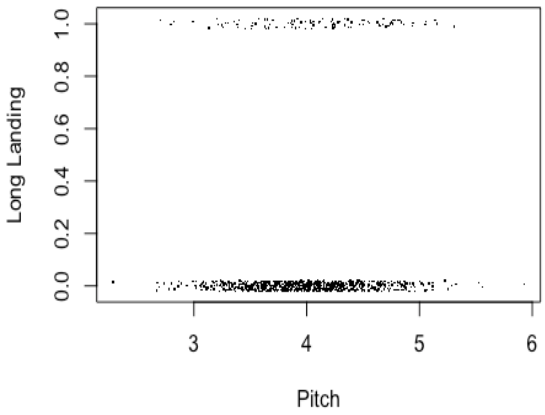


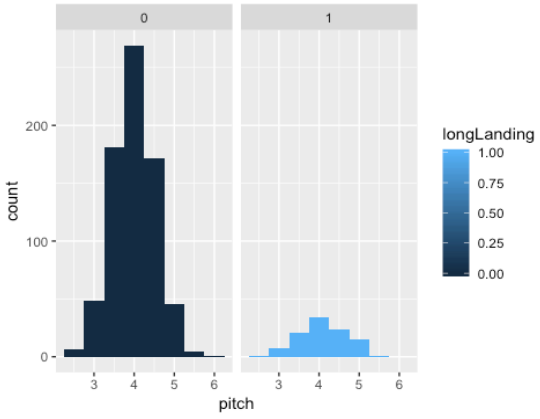


Observations: You can see that there are more flights proportionally that have long landings in the Boeing aircrafts than Airbus.

```
plot(jitter(longLanding, 0.1)~jitter(pitch), FAA, xlab = "Pitch", ylab = "Long Landing",
pch = ".")

ggplot(FAA,aes(x=pitch,fill=longLanding))+geom_histogram(position="dodge",binwidth=0.5) +
  facet_grid(~longLanding)
```

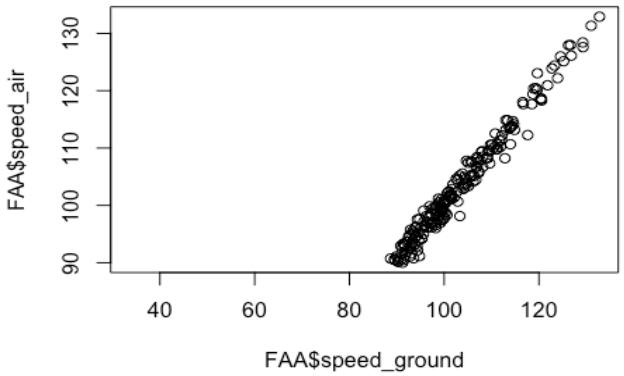




Observations: You can see that the distribution is normal, and that most landings happen at a pitch of 4 regardless if they are long or not.

Step 5

```
plot(FAA$speed_ground, FAA$speed_air)
```




```
proposedModelLong <- glm(longLanding ~ speed_ground + aircraft + pitch, data = FAA,
family = binomial(link = "logit"))
summary(proposedModelLong)

##
## Call:
## glm(formula = longLanding ~ speed_ground + aircraft + pitch,
##     family = binomial(link = "logit"), data = FAA)
##
## 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 ***
## speed_ground     0.61471    0.09184   6.694 2.18e-11 ***
## aircraftboeing   3.04348    0.73345   4.150 3.33e-05 ***
## 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
```

```
exp(coef(proposedModelLong))
```

```
##    (Intercept)   speed_ground aircraftboeing          pitch
##   3.155054e-30   1.849116e+00   2.097815e+01   2.903703e+00
```

```
#making full and empty model for later steps
fullFAA_omitNA <- na.omit(FAA)
fullModel <- glm(longLanding ~ aircraft + no_pasg + speed_ground + speed_air +
                   height + pitch + duration, data = fullFAA_omitNA, family =
binomial(link = "logit"))
summary(fullModel)

emptyModel <- glm(longLanding ~ 1, data = fullFAA_omitNA, family = binomial(link =
"logit"))
summary(emptyModel)
```

Conclusions: Because the variables speed in air and speed on ground are highly correlated to each other, one can and should be eliminated. If you run models with the two variables on their own and then together, you see that the speed in air variable does not change much at all. Where the speed on ground variable does change. Because of this, I am going to proceed with speed on ground instead of speed in air. Another reason to proceed with Speed on Ground, is because there is a large percentage of flights with missing information from the speed in air column, so making a model with a variable with a lot of missing values could skew the model innapropriately. For this reason, I am leaving Speed in Air out of the model, and Speed on Ground in the proposed model. Aircraft and pitch are inlcuded because they are statistically significant on their own (seen in Table 1). The results of the new model can been seen below in Table 2. While Speed on Ground and aircraft are significant, pitch is not significant at alpha=0.05 but it is at alpha=0.10. For now I am going to include pitch in the model.

Table 2.

Rank	Variable	Regression Coefficient	Odd Ratio	Direction of Regression Coefficient	p-value
1	Speed On Ground	0.61471	1.849116	+	2.18e-11
2	Aircraft	3.04348	20.97815	+	3.33e-05
3	Pitch	1.06599	2.903703	+	0.0775

Step 6

```
stepAICModel <- step(proposedModelLong, trace = 0)
summary(stepAICModel)
```

```
##
## Call:
## glm(formula = longLanding ~ speed_ground + aircraft + pitch,
##     family = binomial(link = "logit"), data = FAA)
##
## 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 ***
## speed_ground     0.61471    0.09184   6.694 2.18e-11 ***
## aircraftboeing   3.04348    0.73345   4.150 3.33e-05 ***
## 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
```

```
exp(coef(stepAICModel))
```

```
##    (Intercept)   speed_ground aircraftboeing          pitch
##   3.155054e-30   1.849116e+00   2.097815e+01   2.903703e+00
```

Conclusions: My best model from above (question 5) using the Step AIC function includes variables Aircraft, Speed on Ground, and Pitch. This matches what I predicted from step 3. The results are the exact same as above in Table 2.

Step 7

```
stepBICModel <- step(proposedModelLong, k = log(195), trace = 0)
stepBICModel
```

```
##
## Call:  glm(formula = longLanding ~ speed_ground + aircraft, family = binomial(link =
"logit"),
##     data = FAA)
##
## Coefficients:
##    (Intercept)    speed_ground  aircraftboeing
##       -60.7705          0.5853          3.2368
##
## Degrees of Freedom: 830 Total (i.e. Null);  828 Residual
## Null Deviance:       622.8
## Residual Deviance: 84.66     AIC: 90.66
```

```
exp(coef(stepBICModel))
```

```
##    (Intercept)   speed_ground aircraftboeing
##   4.052395e-27   1.795600e+00   2.545195e+01
```

Conclusions: My best model using the Step BIC function includes variables Aircraft and Speed on Ground. This is partly what I predicted in step 3, but the BIC step variable selection left out pitch, because the BIC penalizes a more complex model, and choose a more simple model. The results of the model are shown below in Table 3.

Table 3.

Rank	Variable	Regression Coefficient	Odd Ratio
1	Speed On Ground	0.5853	1.795600
1	AircraftBoeing	3.2368	25.45195

Step 8

In my opinion, the major risk factors for long landings are the Speed On Ground and the make of the aircraft, with Boeing being more likely to have long landings.

My Model: longLanding = speed_ground + aircraft

This is backed up by a combination of information:

1. The percent of Long Landings is about 12%, which is a significant amount of landings that are getting close to being dangerous.

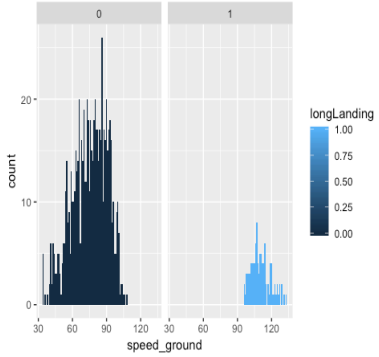
2. The significance of each variable from AIC model selection is significant (except pitch), shown in the table below.

3. I am choosing to leave pitch out of the model because it is not significant at alpha=0.05 and the BIC model selection choose to leave it out to favor a more simple model. I think this is the correct move as well. The summary results of my proposed model are below in Table 4.

Table 4.

Rank	Variable	Regression Coefficient	Odd Ratio	Direction of Regression Coefficient	p-value
1	Speed On Ground	0.58534	1.795600	+	4.08e-12
2	Aircraft	3.23679	25.45195	+	5.45e-06

A visual of speed on ground is shown in the graph to the right as well so you can see that the planes going at higher speeds tend to have long landings, which is backup up statisitically as well.

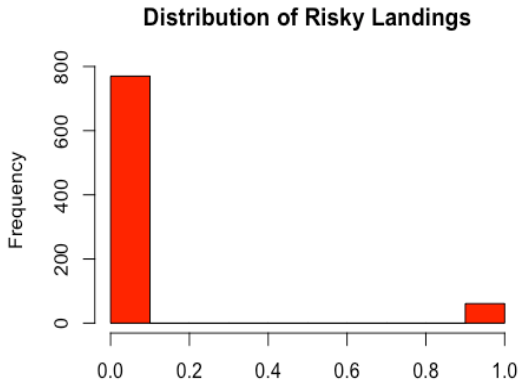


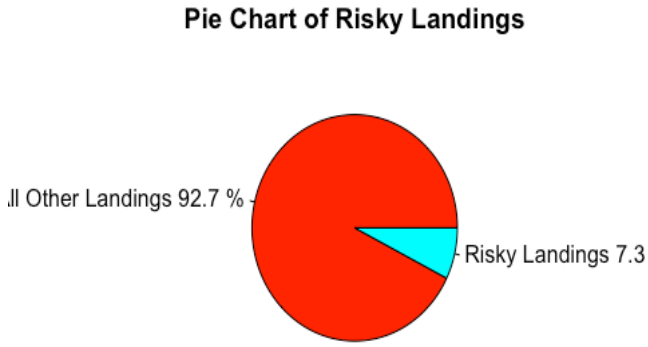
Identify Importnat Factors Using the Binary Data of Risky Landings

Step 9

```
hist(FAA$riskyLanding, main = "Distribution of Risky Landings", col = 'red', xlab = ' ')

pct <- round(table(FAA$riskyLanding) / length(FAA$riskyLanding) * 100, 1)
labs <- c("All Other Landings", "Risky Landings")
labs <- paste(labs, pct)
labs <- paste(labs, "%", sep = " ")
pie(table(FAA$riskyLanding), labels = labs, col = rainbow(length(labs)), main = "Pie
Chart of Risky Landings")
```





Conclusion: There are by far more landings that are not risky However, even if the number of risky landings is infrequent, they are still very important to evaluate. The percentage of risky landings (7.3%) are less than the long landings (12%) but as their name suggests these flights are very risky.

```
#model for no-pasg
modelNoPasgR <- glm(FAA$riskyLanding ~ FAA$no_pasg, family = binomial)
summary(modelNoPasgR)

#model for speed on ground
modelSpeedOnGroundR <- glm(FAA$riskyLanding ~ FAA$speed_ground, family = binomial)
summary(modelSpeedOnGroundR)

#model for speed in air
modelSpeedInAirR <- glm(FAA$riskyLanding ~ FAA$speed_air, family = binomial)
summary(modelSpeedInAirR)

#model for height
modelHeightR <- glm(FAA$riskyLanding ~ FAA$height,family = binomial)
summary(modelHeightR)

#model for pitch
modelPitchR <- glm(FAA$riskyLanding ~ FAA$pitch, family = binomial)
summary(modelPitchR)

#model for aircraft
dummy <- as.numeric(FAA$aircraft == "airbus", 0, 1)

modelAircraftR <- lm(FAA$riskyLanding ~ dummy, data = FAA)
summary(modelAircraftR)
```

```r
#model for duration
modelDurationR <- lm(FAA$riskyLanding ~ duration, data = FAA)
summary(modelDurationR)

##
## Call:
## lm(formula = FAA$riskyLanding ~ duration, data = FAA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08600 -0.07943 -0.07642 -0.07320  0.93313
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0894085  0.0320152   2.793  0.00536 **
## duration    -0.0000813  0.0001975  -0.412  0.68062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2666 on 779 degrees of freedom
##   (50 observations deleted due to missingness)
## Multiple R-squared:  0.0002176,  Adjusted R-squared:  -0.001066
## F-statistic: 0.1696 on 1 and 779 DF,  p-value: 0.6806

#Sample output for all variables
```

Calculating the odds ratio

```r
#calculating the odds ratio
exp(modelNoPasgR$coefficients)
exp(modelSpeedOnGroundR$coefficients)
exp(modelSpeedInAirR$coefficients)
exp(modelHeightR$coefficients)
exp(modelPitchR$coefficients)
exp(modelDurationR$coefficients)
exp(modelAircraftR$coefficients)
```

```r
exp(modelAircraftR$coefficients)

## (Intercept)       dummy
##   1.1146352   0.9363796

#Sample output for all variables
```

Table that Ranks the factors from most important to the least

Table 5.

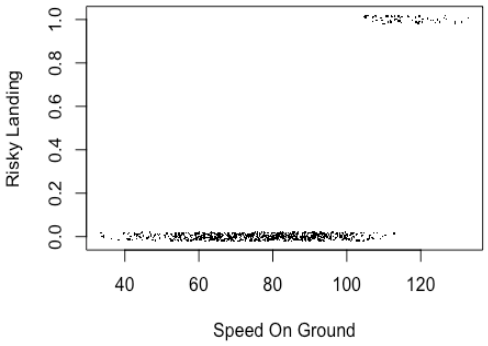
Rank	Variable	Regression Coefficient	Odd Ratio	Direction of Regression Coefficient	p-value
1	Speed On Ground	0.6142	1.848212	+	6.9e-08
2	Speed In Air	0.8704	2.387870	+	3.73e-06
3	Aircraft	-0.06573	0.9363796	-	0.00028
4	Pitch	0.3711	1.44928737	+	0.143296
5	Number of	-0.02538	0.9749400	-	0.154

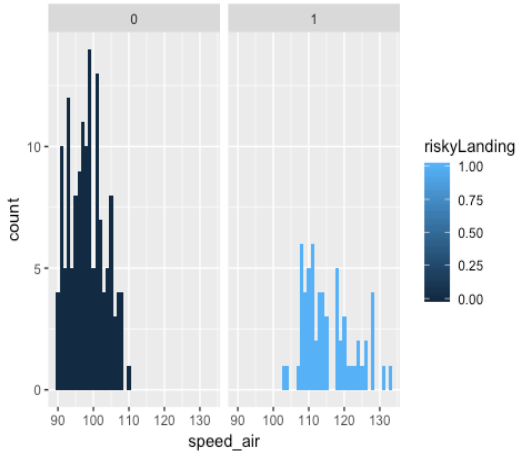
	Passengers				
6	Duration	-0.0000813	0.9999187	-	0.68062
7	Height	-0.002219	0.99778385	-	0.871

Conclusion: From the information above it appears that speed on ground is the most important factor in having a long landing. Followed by Speed In Air and Aircraft, with the rest not looking significant.

```
plot(jitter(riskyLanding,0.1)~jitter(speed_ground), FAA, xlab="Speed On Ground", ylab =
"Risky Landing", pch = ".")

ggplot(FAA, aes(x = speed_ground, fill = riskyLanding)) + geom_histogram(position =
"dodge",binwidth = 1) +
  facet_grid(~riskyLanding)
```

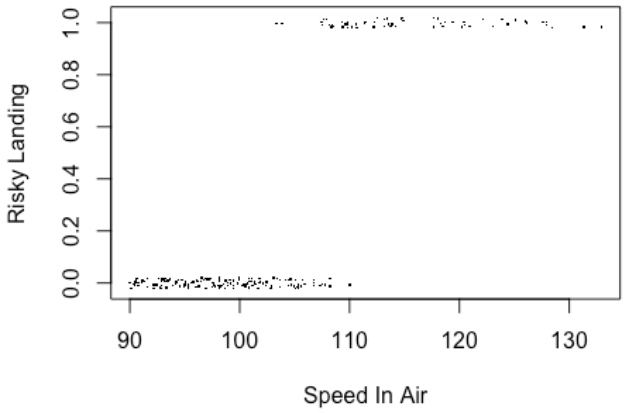


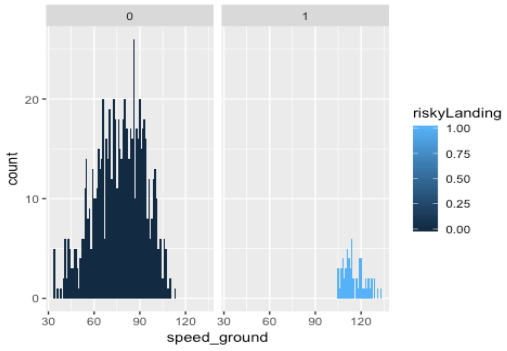


Observations: From the above graphs for 'Speed On Ground' you can see that there is a pattern between Risky Landing and Speed on Ground. Most of the flights that had risky landings were going at a faster speed on ground than their counterparts. Also, the distribution of speed on ground is normal.

```
plot(jitter(riskyLanding, 0.1)~jitter(speed_air), FAA, xlab = "Speed In Air", ylab =
"Risky Landing", pch = ".")

ggplot(FAA, aes(x = speed_air, fill = riskyLanding)) + geom_histogram(position =
"dodge",binwidth = 1) +
  facet_grid(~riskyLanding)
```



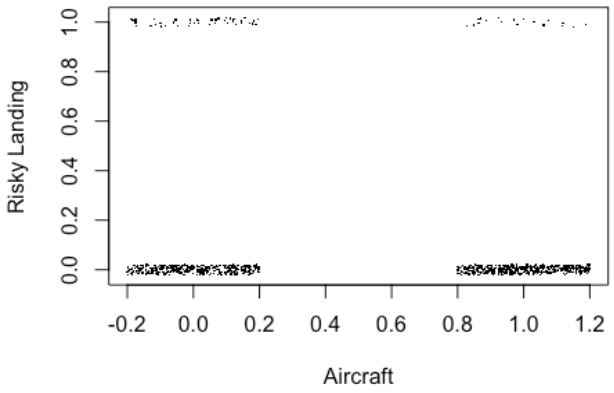


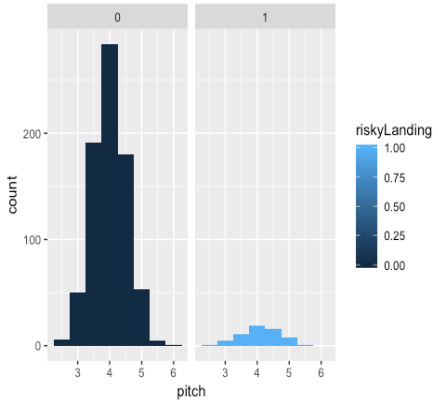
Observations: You can see that there is a pattern between Risky Landing and Speed in Air. Most of the flights that had risky landings were going at a faster speed on ground than their counterparts, however

there still are some flights that are going slower that are still going over the runway. Also, the distribution of speed on ground is realatively normal.

```
plot(jitter(riskyLanding, 0.1)~jitter(dummy), FAA, xlab = "Aircraft", ylab = "Risky
Landing", pch = ".")

ggplot(FAA, aes(x = riskyLanding, fill = aircraft)) +
  geom_bar(position = "dodge", width = 0.5) +
  facet_grid(~aircraft)
```

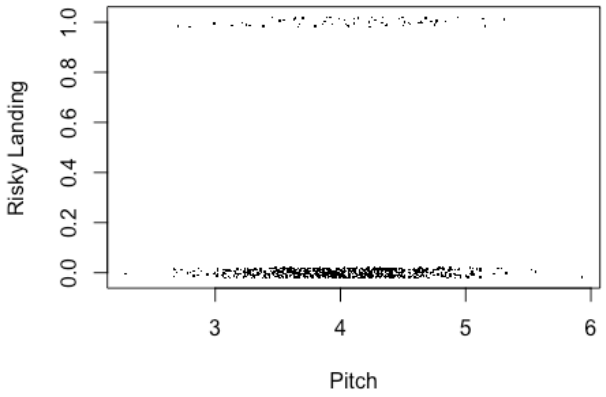


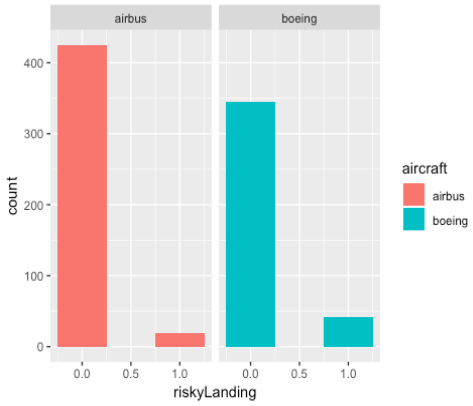


Observations: You can see that there are more flights that have risky landings in the Boeing aircrafts than Airbus.

```
plot(jitter(riskyLanding, 0.1)~jitter(pitch), FAA, xlab = "Pitch", ylab = "Risky
Landing", pch = ".")

ggplot(FAA,aes(x=pitch,fill =
riskyLanding))+geom_histogram(position="dodge",binwidth=0.5) +
  facet_grid(~riskyLanding)
```





Observations: You can see that the distribution is normal, and that most landings happen at a pitch of 4 regardless if they are risky or not.

```r
#finding the model for risky
proposedModelRisky <- glm(riskyLanding ~ speed_ground + aircraft, data = FAA, family =
binomial(link = "logit"))
summary(proposedModelRisky)
```

```
##
## Call:
## glm(formula = riskyLanding ~ speed_ground + aircraft, family = binomial(link =
"logit"),
##     data = FAA)
##
## 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
```

```r
exp(coef(proposedModelRisky))
```

```
##    (Intercept)   speed_ground aircraftboeing
##   4.660321e-45   2.525084e+00   5.564717e+01
```

```r
#these steps performed for something late in the code
fullFAA_omitNA <- na.omit(FAA)
fullModelRisky <- glm(riskyLanding ~ aircraft + no_pasg + speed_ground + speed_air +
                        height + pitch + duration, data = fullFAA_omitNA, family =
binomial(link = "logit"))
summary(fullModelRisky)

emptyModel <- glm(longLanding ~ 1, data = fullFAA_omitNA, family = binomial(link =
"logit"))
summary(emptyModel)
```

Conclusions: Because the variables speed in air and speed on ground are highly correlated to each other, one can and should be eliminated. If you run models with the two variables on their own and then together, you see that the speed in air variable does not change much at all. Where the speed on ground variable does change. Because of this, I am going to proceed with speed on ground instead of speed in air. Also because there are many missing observations for speed in air, so preceeding with that variable could mess with the model inapropriately. I am picking speed on ground for the risky landing model, same as the long landings.

```
stepAICModel <- step(proposedModelRisky, trace = 0)
stepAICModel

##
## Call:  glm(formula = riskyLanding ~ speed_ground + aircraft, family = binomial(link =
"logit"),
##     data = FAA)
##
## Coefficients:
##    (Intercept)    speed_ground  aircraftboeing
##      -102.0772          0.9263          4.0190
##
## Degrees of Freedom: 830 Total (i.e. Null);  828 Residual
## Null Deviance:       436
## Residual Deviance: 40.1  AIC: 46.1
```

Conclusions: My best model using the Step AIC function includes variables Aircraft and Speed in Air. This matches what I predicted from step 3.

```
stepBICModel <- step(fullModelRisky, k = log(195), trace = 0)
stepBICModel

##
## Call:  glm(formula = riskyLanding ~ aircraft + speed_air, family = binomial(link =
"logit"),
##     data = fullFAA_omitNA)
##
## Coefficients:
##    (Intercept)  aircraftboeing       speed_air
##       -133.717           4.550           1.221
##
## Degrees of Freedom: 194 Total (i.e. Null);  192 Residual
## Null Deviance:       240.7
## Residual Deviance: 26.28     AIC: 32.28

exp(coef(stepBICModel))

##    (Intercept) aircraftboeing      speed_air
##   8.457903e-59   9.462423e+01   3.389270e+00
```

Conclusions: My best model using the Step BIC function includes variables Aircraft and Speed on Ground. This also matches what I predicted from step 3. They also match each other, the AIC and BIC selection tool get the same model. Which is great! Indicates this model is a good fit.

Step 10

The two major risk factors for Risky landings are the Speed In Air and the make of the aircraft. If you can focus on these two variables, potentially many unsafe landings can be avoided.

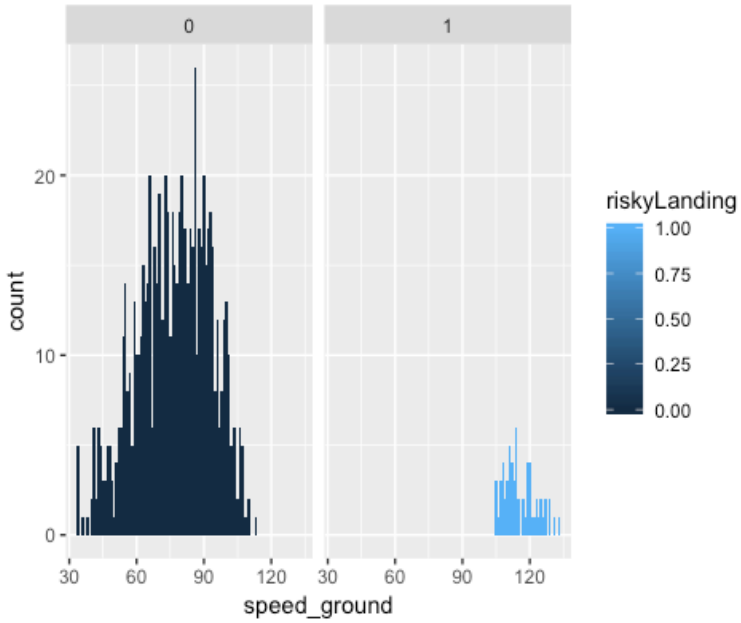
My Model: RiskyLanding = speed_ground + aircraft

This is backed up by a combination of valuable information:
1. About 7.3% of all flights are considered “Risky” landings, that is a landing that goes over 3000 feet, this is a very high percentage.

2. The significance of each variable on Risky Landing is significant, shown in the table below.
3. The following visualization of the speed on ground variable, you can see the aircrafts that are traveling

at a higher speed on ground have more risky landing distances.
4. I am choosing to leave pitch out of the model, as the BIC and AIC step forward variable selection did not include it, and it is not significant if it is included in the model.

Rank	Variable	Regression Coefficient	Odd Ratio	Direction of Regression Coefficient	p-value
1	Speed On Ground	0.9263	2.525084	+	3.78e-05
2	Aircraft	4.0190	55.64717	+	0.0013



Compare the Two Models Built for Long Landing and Risky Landing

Step 11

The similarities and differences are summarized below:

- The two models are similar; however the Risky Landings drop the variable pitch as factors impacting risky landing. This could mean that pitch is important in predicting a long landing, but not necessarily a risky one.
- Speed on ground for risky flights has a larger odds ratio, meaning that for the risky flights if the speed on the ground increases by 1 mph, then the flight is 2.5 more likely to have a risky landing than a regular landing. Compared to the long landings which has an odds ratio of 1.8, which still large, just not as big of an impact as the risky landing.
- The difference in the odds ratio makes sense also, because if the speed on the ground goes up by 1mph then the risky landing is impacted by 2.5, vs the long landing of 1.8. When speed goes up then the changes of having a risky landing are higher.

Step 12

```
#ROC for Long model
linPred <- predict(proposedModelLong)
predProb <- predict(proposedModelLong, type = "response")
predOut <- ifelse(predProb < 0.05, "no", "yes")
```

```
fullFAA_omitNA <- data.frame(FAA, predProb, predOut)
xtabs(~longLanding+predOut, FAA)
```

```
##             predOut
## longLanding  no yes
##           0 681  47
##           1   0 103
```

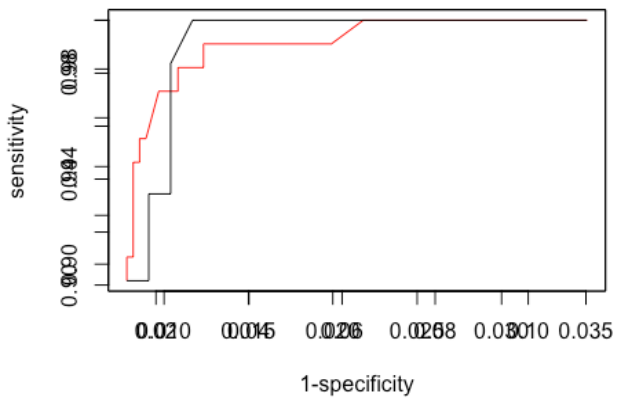
```
thresh<-seq(0.01,0.5,0.01)
sensitivity<-specificity<-rep(NA,length(thresh))
for(j in seq(along=thresh)){
  pp<-ifelse(fullFAA_omitNA$predProb<thresh[j],"no","yes")
  xx<-xtabs(~longLanding+pp,fullFAA_omitNA)
  specificity[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}

#ROC for Risky Model
linPredR <- predict(proposedModelRisky)
predProbR <- predict(proposedModelRisky, type = "response")
predOutR <- ifelse(predProbR < 0.05, "no", "yes")
fullFAA_omitNA <- data.frame(FAA, predProbR, predOutR)
xtabs(~riskyLanding + predOutR, FAA)
```

```
##             predOutR
## riskyLanding  no yes
##            0 755  15
##            1   0  61
```

```
thresh<-seq(0.01,0.5,0.01)
sensitivity1<-specificity1<-rep(NA,length(thresh))
for(j in seq(along=thresh)){
  pp<-ifelse(fullFAA_omitNA$predProbR<thresh[j],"no","yes")
  xx<-xtabs(~riskyLanding+pp,fullFAA_omitNA)
  specificity1[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity1[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}

#plot both together
plot(1-specificity,sensitivity,type="l", col = "red", xlab = "1-specificity", ylab =
"sensitivity")
par(new = TRUE)
plot(1-specificity1,sensitivity1,type="l", xlab = " ", ylab = " ")
```



Conclusion: In the above figure, the red line is the performance of the long landing model, and the black is the performance of the risky landing model. The performance of both are very similar with the area under curve. However, I would say the performance of the long landing model is slightly better, which makes sense because the long landing model encompases the risky landings as well. They both do look pretty good though, which is great.

Step 13

We will now use our models to predict the landings of this new flight that is coming in to land at an airport. The following line has the data for the new flight.

```
#new flight with new information
newFlight <- data.frame(aircraft = "boeing", duration = 200, no_pasg = 80, speed_ground =
115, speed_air = 120, height = 40, pitch = 4)

#predicting if the model will be a long landing
predictModelLong <- predict(proposedModelLong, newdata = newFlight, type = 'response' ,
se = T)
c(predictModelLong$fit, predictModelLong$fit-1.96*predictModelLong$se.fit[1],
predictModelLong$fit+1.96*predictModelLong$se.fit[1])

##         1         1         1
## 0.9999577 0.9998236 1.0000917

#predicting if the model will be a risky landing
predictModelRisky <- predict(proposedModelRisky, newdata = newFlight, type = "response",
se = TRUE)
c(predictModelRisky$fit, predictModelRisky$fit-1.96*predictModelRisky$se.fit[1],
predictModelRisky$fit+1.96*predictModelRisky$se.fit[1])

##        1        1        1
## 0.999789 0.998925 1.000653
```

Conclusion: Now that we have models that are performing well, we can use a new flight with data to predict whether it will be long or risky. The new flight has a probabillity of 0.9999577 of being a long landing, at a 95% confidence interval of (0.9998236, 1.0000917). The new flight has a probability of 0.999789 of being a risky landing, at a 95% confidence interval of (0.998925, 1.000653). The model indicates it has a very high probability of landing long and risky. This is consistent with what is expected, however this is incredibly high, and nothing should have a prediction over 1.0.

Compare models with different link functions

Step 14

We now will look at our risky landing model with the three different linkers, logit, probit, and cLogLog. to evaluate which has the best performance, evaluated with the AIC score.

```
ModelRiskyLogit <- glm(riskyLanding ~ speed_ground + aircraft, data = FAA, family =
binomial(link = "logit"))

ModelRiskyProbit <- glm(riskyLanding ~ speed_ground + aircraft, data = FAA, family =
binomial(link = probit))

ModelRiskyLogLog <- glm(riskyLanding ~ speed_ground + aircraft, data = FAA, family =
binomial(link = cloglog))
```

```
round(coef(ModelRiskyLogit),3)
round(coef(ModelRiskyProbit),3)
round(coef(ModelRiskyLogLog),3)

summary(ModelRiskyLogit)
summary(ModelRiskyProbit)
summary(ModelRiskyLogLog)
```

Conclusion: Given the results in the table below, especially the AIC values, the best link to use would be the Probit linker, because it results in the lowest AIC value. However, all three are incredibly close. All three models are good to pick, in reality the different linkers do not make a huge difference.

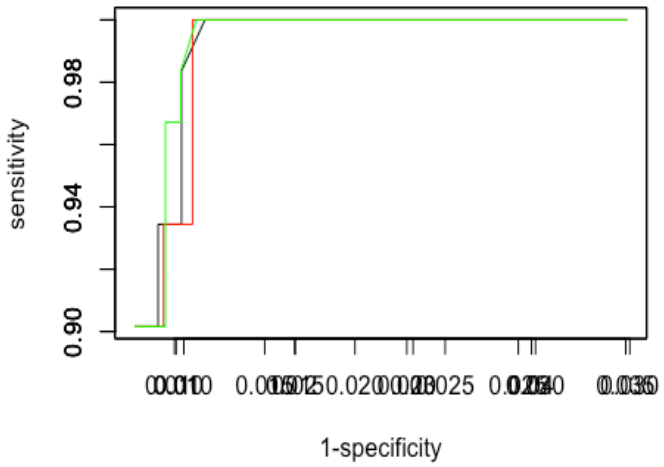
	Logit Model	Probit	cLogLog
Aircraft Est	4.019	2.357	2.898
Speed Ground Est	0.926	0.532	0.622
AIC	46.097	45.436	47.443

Step 15

```
#logit model
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(ModelRiskyLogit,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~FAA$riskyLanding+pp)
  specificity_r[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity_r[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}
#probit model
thresh <- seq(0.01,0.5,0.01)
sensitivity_probit <- specificity_probit <- rep(NA,length(thresh))
for( j in seq(along=thresh)) {
  pp<- ifelse(predict(ModelRiskyProbit,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~FAA$riskyLanding+pp)
  specificity_probit[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity_probit[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}
#cloglog model
thresh <- seq(0.01,0.5,0.01)
sensitivity_loglog <- specificity_loglog <- rep(NA,length(thresh))
for( j in seq(along=thresh)) {
  pp<- ifelse(predict(ModelRiskyLogLog,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~FAA$riskyLanding+pp)
  specificity_loglog[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity_loglog[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}

#plotting all three together
plot(1 - specificity_r, sensitivity_r, type = "l", col = "black", xlab = " ", ylab =
"sensitivity")
par(new = TRUE)
plot(1 - specificity_probit, sensitivity_probit, type = "l", col = "red", xlab = " ",
ylab = " ")
par(new = TRUE)
```

```
plot(1 - specificity_loglog, sensitivity_loglog, type = "l", col = "green", xlab = "1-
specificity", ylab = " ")
```



Conclusion: In the figure above, the black curve is the logit linker, red is probit, and green is the cLogLog linker. The ROC curves show the performance of the three models is very similar. The AUC looks to be the same in all three model linkers. Again showing that the three linkers do not have a huge impact on performance.

Step 16

We will use the below code to identify the top 5 risky landings in each model.

```
predRiskyLogit <- predict(ModelRiskyLogit,type = 'response')
predRiskyProbit <- predict(ModelRiskyProbit,type = 'response')
predRiskyLoglog <- predict(ModelRiskyLogLog,type = 'response')

#sort by tail, so you get the obs with high risk
FAA[as.numeric(names(tail(sort(predRiskyLogit),5))),]
FAA[as.numeric(names(tail(sort(predRiskyProbit),5))),]
FAA[as.numeric(names(tail(sort(predRiskyLoglog),5))),]
```

Conclusion: Probit and cLogLog have the most similar risky landings. Probits top 5 risky landings are: 675, 772, 773, 784, 814 and cLogLog is: 760, 772, 773, 784, and 814. These two have all the same top flights, except for their first one. However, the logit linker gives much different flights as being the most risky. Their flights are 227, 675, 814, 773, 513, there are a few similar, but not as much similarity as the other two. These flights could be used to improve the models as flights that would be extremely risky, and good flights for FAA agents to be on the lookout for in advising against landing the plane.

Step 17

```
#make new models with linkers in variable name to not get confused
proposedModelRiskyLogit <- glm(riskyLanding ~ speed_ground + aircraft, data = FAA, family
= binomial(link = "logit"))
summary(proposedModelRiskyLogit)

proposedModelRiskyProbit <- glm(riskyLanding ~ speed_ground + aircraft, data = FAA,
family = binomial(link = "probit"))
```

```r
proposedModelRiskyCLoglog <- glm(riskyLanding ~ speed_ground + aircraft, data = FAA,
family = binomial(link = "cloglog"))

#straight from notes in class
predLogitRisky <- predict(proposedModelRiskyLogit, newdata = newFlight, type =
'response', se.fit = T)
c(predLogitRisky$fit, predLogitRisky$fit - 1.96 * predLogitRisky$se.fit[1],
predLogitRisky$fit + 1.96 * predLogitRisky$se.fit[1])

##        1        1        1
## 0.999789 0.998925 1.000653

predProbitRisky <- predict(proposedModelRiskyProbit, newdata = newFlight, type =
'response', se.fit = T)
c(predProbitRisky$fit, predProbitRisky$fit - 1.96 * predProbitRisky$se.fit[1],
predProbitRisky$fit + 1.96 * predProbitRisky$se.fit[1])

##         1         1         1
## 0.9999994 0.9999933 1.0000056

predCLoglogRisky <- predict(proposedModelRiskyCLoglog, newdata = newFlight, type =
'response', se.fit = T)
c(predCLoglogRisky$fit, predCLoglogRisky$fit - 1.96 * predCLoglogRisky$se.fit[1],
predCLoglogRisky$fit + 1.96 * predCLoglogRisky$se.fit[1])

## 1 1 1
## 1 1 1
```

Conclusion: The different linkers performance is also very similar at predicting a risky landing. The logit linker can predict a risky flight landing at a probabillity of 0.9999577, at a 95% confidence interval of (0.9998236, 1.0000917). The probit linker predicts a risky landing at a probability of 0.9999994, at a 95% cofidence interval of (0.9999933, 1.0000056), and the hazard linker (cLogLog) predicts a risky landing at a probability of 1, at a 95% cofidence interval of (1,1). All three linkers can predict at a very high probability which indicates that perhaps all three models, irrespective of the linker used, are overfit to the data. This could have potentially been avoided by using a training set of data and a testing set, instead of building the model on 100% of the data.