Stat_Modelling_Hw_2

Digvijay Kawale

2/4/2020

Project Part 2: Study of Logistic Regression

Loading the required Packages

```
library(tidyverse)

## ── Attaching packages ──────────────────────────────────────────────────────────── tidyverse 1.2.1
──

## ✔ ggplot2 3.2.1     ✔ purrr   0.3.2
## ✔ tibble  2.1.3     ✔ dplyr   0.8.3
## ✔ tidyr   0.8.3     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0

## ── Conflicts ──────────────────────────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()

library(readxl)
library(dplyr)
library(corrr)
library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

Step 0: Getting the cleaned data set from steps 1 to 9 of Project Part 1.

- The final data set that we used for the study of linear regression had 832 observations of 8 variables. The same data set is obtained using the steps followed in the project part 1.Those steps are outlined as comments in the below code part.

```
## Loading Data Sets

Flights_800 <- read_xls("~/Desktop/Subjects/Flex 3/Statistical Modelling/WeeK
1/FAA1-1.xls")
Flights_150 <- read_xls("~/Desktop/Subjects/Flex 3/Statistical Modelling/WeeK
1/FAA2-1.xls")

## Merging Two Data Sets and removing duplicates

Flights_150$duration <- NA

flights_final <- rbind(Flights_800, Flights_150)

flights_columns <- flights_final[c("aircraft"   ,  "no_pasg"   ,
"speed_ground" ,"speed_air"   , "height"   ,    "pitch"    , "distance"   )]

flights_final <- flights_final[!duplicated(flights_columns),]

## Removing abnormal observations from the data set

flights_final <- filter(flights_final, ifelse(is.na(height), TRUE, height >=
6))

flights_final <- filter(flights_final, ifelse(is.na(speed_ground), TRUE,
(speed_ground >= 30 & speed_ground <= 140)))

flights_final <- filter(flights_final, ifelse(is.na(speed_air), TRUE,
(speed_air >= 30 & speed_air <= 140)))

flights_final <- filter(flights_final, ifelse(is.na(duration), TRUE, duration
>= 40 ))

dim(flights_final)

## [1] 832   8
```

Creating Binary Responses

Step 1: Cretaing the Binary Variables 'long_landing', 'risky_landing' and removing the continous variable for 'distance'.

- A binary response of long landing is created based on the varible distance. If the distance is greater than 2500 then variable long landing will be 1 else it will be 0.
- A binary response of risky landing is created based on the varible distance. If the distance is greater than 3000 then variable risky landing will be 1 else it will be 0.

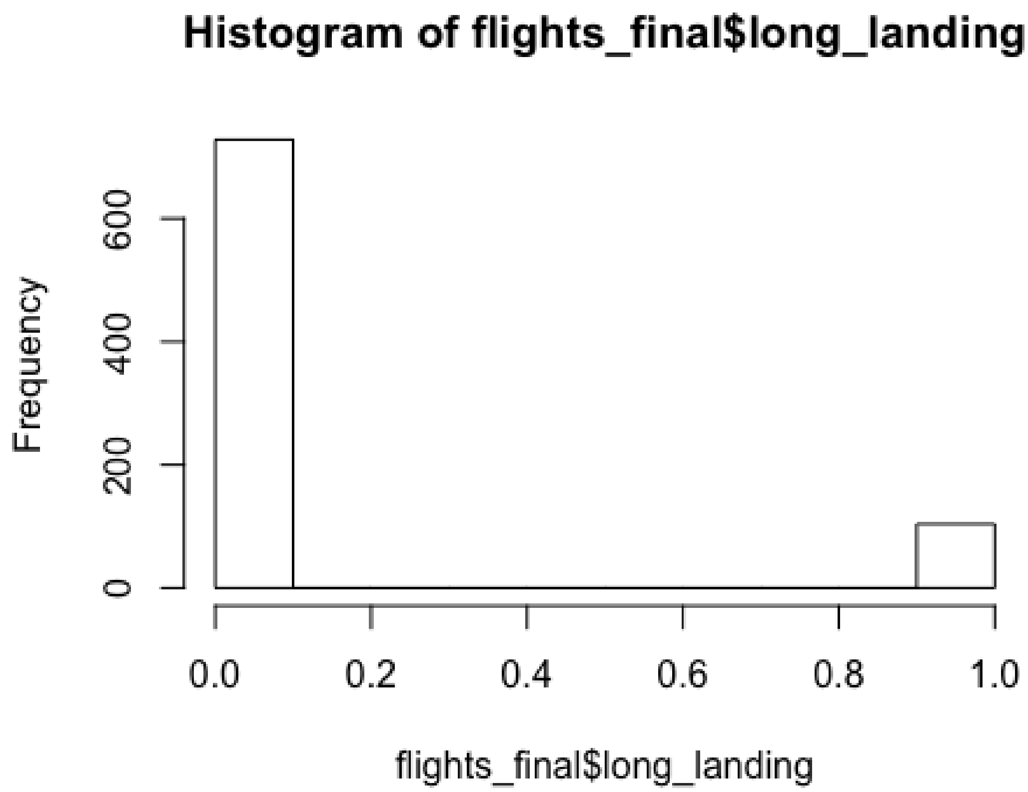
```
## Adding Binary Variables

flights_final$long_landing <- ifelse(flights_final$distance > 2500, 1, 0)

flights_final$risky_landing <- ifelse(flights_final$distance > 3000, 1, 0)

## Discarding the Continous variable 'distance'

flights_final$distance <- NULL
```

Identifying important factors using the binary data of "long_landing"

Step 2: Histogram showing the distribution of long_landing

- It is observed that 104 observations of long_landing have the value 1 and the rest 728 have the value 0.

```
hist(flights_final$long_landing)
```



Step 3: Fitting single-factor logistic regression

- The variable Long Landing is logistically regressed with all the variables one by one.
- Later the results of all the regression models are tabulated in table 1 that contains the size regression coefficient, direction of coeficient, odds ratio and p values.
- Using p- values from the table 1, it is observed that the significant predictor variables are speed_air, speed_ground, pitch and aircraft_num.

```
## Converting the variable 'aircraft' into binary

flights_final$aircraft_num <-  ifelse(flights_final$aircraft == "airbus", 1,
0)

## Fitting single-factor logistic regression using each variable

duration <-  glm(long_landing ~ duration, family = binomial, data =
flights_final)
no_pasg <- glm(long_landing ~ no_pasg, family = binomial, data =
flights_final)
speed_ground <- glm(long_landing ~ speed_ground, family = binomial, data =
flights_final)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

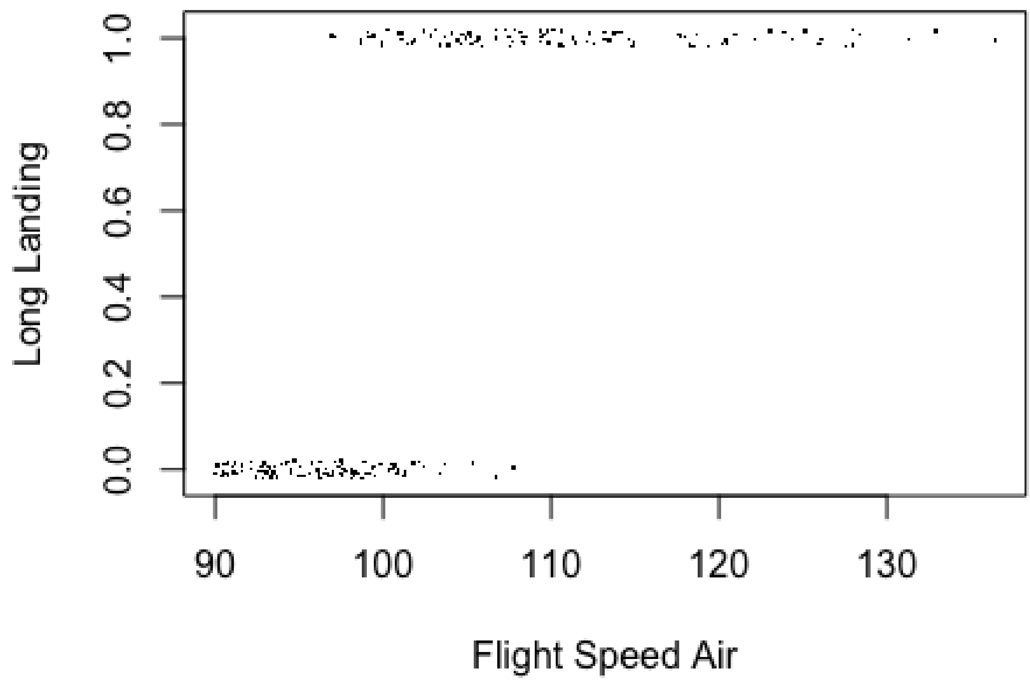
```
speed_air <- glm(long_landing ~ speed_air, family = binomial, data =
flights_final)
height <- glm(long_landing ~ height, family = binomial, data = flights_final)
pitch <- glm(long_landing ~ pitch, family = binomial, data = flights_final)
aircraft_num <- glm(long_landing ~ aircraft_num, family = binomial, data =
flights_final)


##Calculating odds ratio

odds_ratio <- c(
exp(summary(duration)$coefficients[2,1]),
exp(summary(no_pasg)$coefficients[2,1]),
exp(summary(speed_ground)$coefficients[2,1]),
exp(summary(speed_air)$coefficients[2,1]),
exp(summary(height)$coefficients[2,1]),
exp(summary(pitch)$coefficients[2,1]),
exp(summary(aircraft_num)$coefficients[2,1]))


## Creating Variable names vector

variable_names <- c("duration", "no_pasg", "speed_ground", "speed_air",
"height", "pitch", "aircraft_num")


## P values

p_values <- c(
summary(duration)$coefficients[2,4],
summary(no_pasg)$coefficients[2,4],
summary(speed_ground)$coefficients[2,4],
summary(speed_air)$coefficients[2,4],
summary(height)$coefficients[2,4],
summary(pitch)$coefficients[2,4],
summary(aircraft_num)$coefficients[2,4])

## Regression Coefficients

regression_coefficients <- c(
summary(duration)$coefficients[2,1],
summary(no_pasg)$coefficients[2,1],
summary(speed_ground)$coefficients[2,1],
summary(speed_air)$coefficients[2,1],
summary(height)$coefficients[2,1],
summary(pitch)$coefficients[2,1],
summary(aircraft_num)$coefficients[2,1])
```

```
Table_1 <- data.frame(variable_names, regression_coefficients, odds_ratio,
coef_direction = ifelse(regression_coefficients < 0, "Negative", "Positive")
, p_values)

Table_1

##   variable_names regression_coefficients odds_ratio coef_direction
## 1       duration            -0.001211113  0.9987896       Negative
## 2        no_pasg            -0.006523928  0.9934973       Negative
## 3   speed_ground             0.472345761  1.6037518       Positive
## 4      speed_air             0.512321769  1.6691621       Positive
## 5         height             0.009923535  1.0099729       Positive
## 6          pitch             0.403385772  1.4968842       Positive
## 7   aircraft_num            -0.878934945  0.4152249       Negative
##       p_values
## 1 5.850450e-01
## 2 6.414223e-01
## 3 3.935303e-14
## 4 4.333606e-11
## 5 3.530158e-01
## 6 4.436662e-02
## 7 6.090379e-05
```

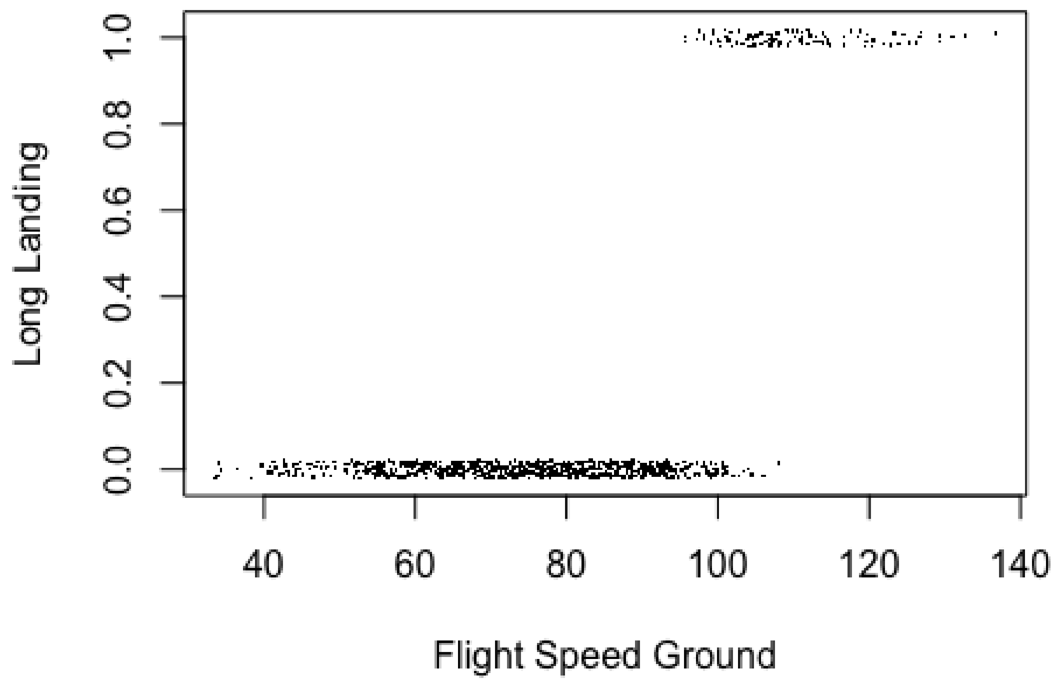
Step 4 : Seeing the association of long landing

- The significance of variables is checked using the p values. The models having p-values less than 0.05 are considered as significant.
- The significant predictor variables observed in table_1 are speed_air, speed_ground, pitch and aircraft_num.

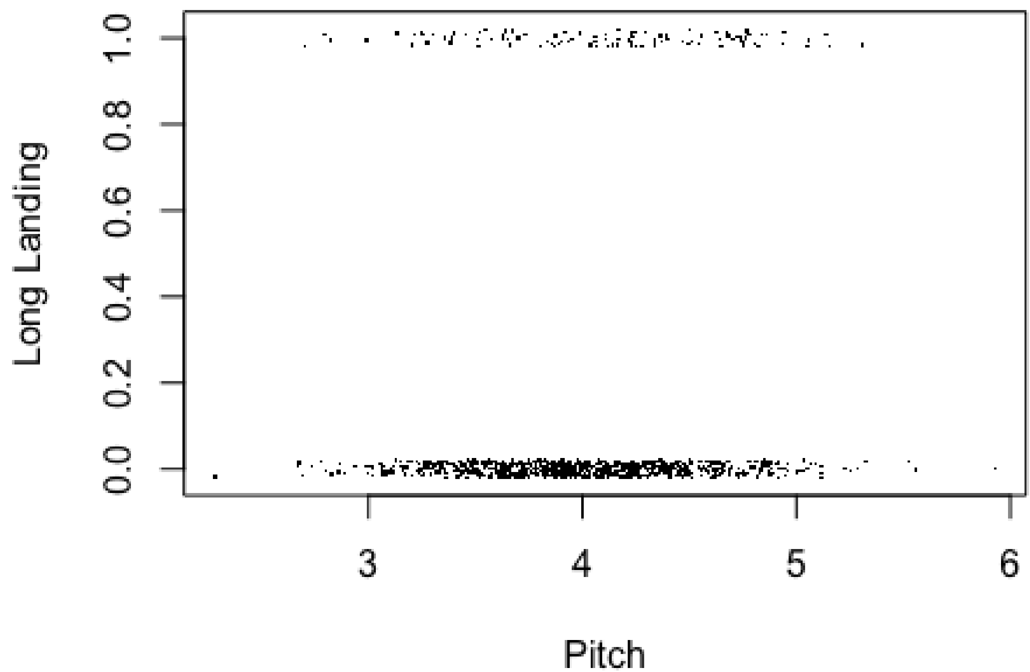
```
##Speed_air

plot(jitter(long_landing,0.1) ~ jitter(speed_air), flights_final, xlab =
"Flight Speed Air", ylab = "Long Landing", pch = ".")
```



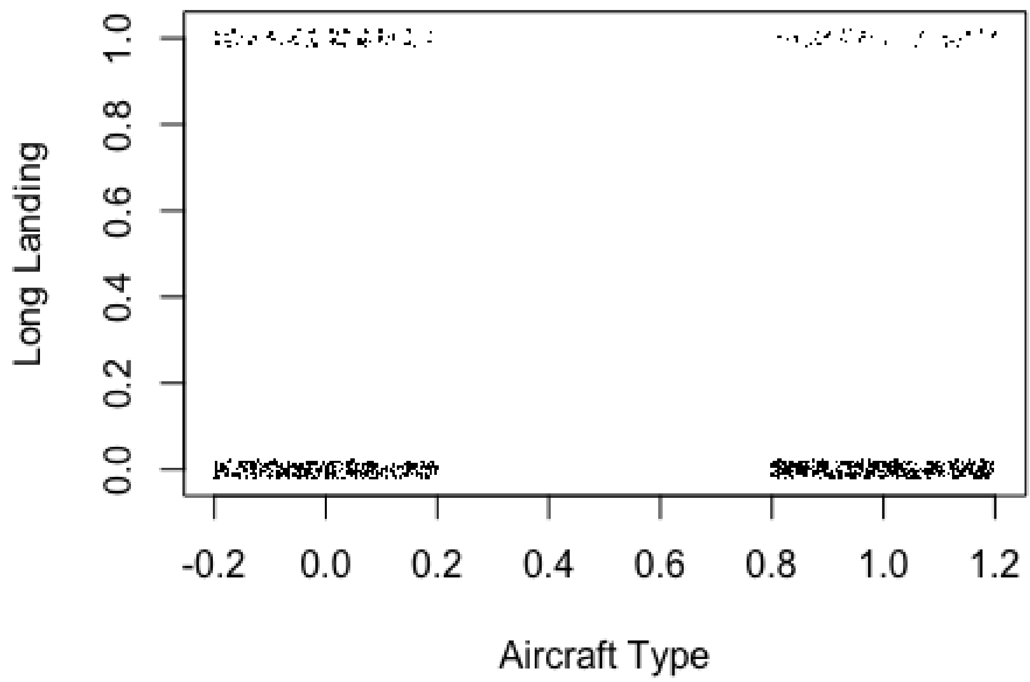

```
##Speed_ground

plot(jitter(long_landing,0.1) ~ jitter(speed_ground), flights_final, xlab =
"Flight Speed Ground", ylab = "Long Landing", pch = ".")
```




```
##Pitch

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




```r
##Aircraft Numeric

plot(jitter(long_landing,0.1) ~ jitter(aircraft_num), flights_final, xlab =
"Aircraft Type", ylab = "Long Landing", pch = ".")
```



Step 5: Fitting the data with all variables together

- It ws observed in step 16 of Project part 1 that the speed air and speed ground were highly collinear. We used speed ground as predictor because the number of NA's in data were high for speed air. Also, speed ground was more significant than speeed air.
- We will now fit a logistic regression using three variables together. The varibles that we will use are speed_ground, pitch and aircraft numeric.
- The full model logistetic regression model tells us that wih a unit increase in Speed Ground the odds ratio will increase by 1.849 when all other variables are kept constant.
- The full model logistetic regression model tells us that wih a unit increase in Pitch the odds ratio will increase by 2.9 when all other variables are kept constant.
- The full model logistetic regression model tells us that wih a unit increase in Aircraft Numeric the odds ratio will increase by 0.047 when all other variables are kept constant.

```
full_model <-  glm(long_landing ~ speed_ground + pitch + aircraft_num, family
= binomial, data = flights_final)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
## Calculating odds ratio

odds_ratio_full_model <- c(
exp(summary(full_model)$coefficients[2,1]),
exp(summary(full_model)$coefficients[3,1]),
exp(summary(full_model)$coefficients[4,1]))

summary(full_model)

##
## Call:
## glm(formula = long_landing ~ speed_ground + pitch + aircraft_num,
##     family = binomial, data = flights_final)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.11589  -0.01114  -0.00026   0.00000   2.40741
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -64.88507   10.11708  -6.413 1.42e-10 ***
## speed_ground   0.61471    0.09184   6.694 2.18e-11 ***
## pitch          1.06599    0.60389   1.765   0.0775 .
## aircraft_num  -3.04348    0.73345  -4.150 3.33e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 626.946  on 831  degrees of freedom
## Residual deviance:  81.309  on 828  degrees of freedom
## AIC: 89.309
##
## Number of Fisher Scoring iterations: 10
```

Step 6: Step Wise AIC

- We will use the Stepwise AIC funtion in R to do the variable selection for the full model of Logistic Regression.
- Before doing that we will remove the character variable aircraft type from the data frame as we have already coded it as binary. We will also remove the speed air variable as it has a lot of NULL values and it is highly collinear with speed ground.
- After applying the step AIC function to the model, is shows that it has lowest AIC of 63.2 when the variables speed_ground, aircraft_num, pitch and height are used. Also, the AIC for the model with variables speed_ground and aircraft_num is 90.66. Since this difference is not large we choose the latter model. Another reason behind that is we have already seen that height and pitch were not significant in the earlier steps.

```r
## Filtering the character variable aircraft and speed air

flights_1 <- dplyr::select(flights_final, duration, no_pasg, speed_ground,
height, pitch, aircraft_num, long_landing)

GLM_long_landing_null <- glm(long_landing ~ 1, family = binomial, data =
flights_1)
GLM_long_landing_full <- glm(long_landing ~ ., family = binomial, data =
flights_1)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

fit1_GLM <- step(GLM_long_landing_null, scope = list(lower
=GLM_long_landing_null,upper = GLM_long_landing_full),    direction =
'forward')

## Start:  AIC=628.95
## long_landing ~ 1

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##                Df Deviance    AIC
## + speed_ground  1   107.40 136.55
## + aircraft_num  1   586.99 616.14
## + pitch         1   599.11 628.26
## <none>              601.79 628.95
## + height        1   601.21 630.36
## + duration      1   601.50 630.65
## + no_pasg       1   601.60 630.75

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=119.47
## long_landing ~ speed_ground

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##                 Df Deviance     AIC
## + aircraft_num  1   78.164  92.233
## + height        1   95.059 109.129
## + pitch         1   97.006 111.076
## <none>              107.401 119.470
## + duration      1  107.296 121.365
## + no_pasg       1  107.375 121.444

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=90.66
## long_landing ~ speed_ground + aircraft_num

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##             Df Deviance    AIC
## + height     1   54.401 68.902
## + pitch      1   75.176 89.677
## <none>           78.164 90.665
## + duration  1   76.635 91.136
## + no_pasg   1   77.824 92.325

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Step:  AIC=65.05
## long_landing ~ speed_ground + aircraft_num + height

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##             Df Deviance    AIC
## + pitch      1   51.580 64.225
## <none>           54.401 65.047
## + duration  1   53.680 66.325
## + no_pasg   1   54.401 67.047

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=63.2
## long_landing ~ speed_ground + aircraft_num + height + pitch

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##             Df Deviance    AIC
## <none>           51.580 63.204
## + duration  1   51.102 64.726
## + no_pasg   1   51.575 65.199
```

Step 7: Step Wise BIC

- The step function in R can also be used with BIC as our parameter. We will give an extra argument 'k = log(nrow(flights_1))' in the step function. The use of this function for BIC was found through google search. Here is its link
- We observe similar kind of results in stepwise BIC as well. The BIC of 104.84 is observed when the variables speed ground and aircraft numeric are used as predictors.
- Therefore, the final variables that we will be using as predictors are speed ground and aircraft numeric.

```
fit2_GLM <- step(GLM_long_landing_null, scope = list(lower
=GLM_long_landing_null,upper = GLM_long_landing_full),    direction =
'forward', k = log(nrow(flights_1)))
```

```
## Start:  AIC=633.67
## long_landing ~ 1

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                  Df Deviance    AIC
## + speed_ground    1   107.40 146.00
## + aircraft_num    1   586.99 625.59
## <none>                601.79 633.67
## + pitch           1   599.11 637.71
## + height          1   601.21 639.81
## + duration        1   601.50 640.09
## + no_pasg         1   601.60 640.20

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## 
## Step:  AIC=128.92
## long_landing ~ speed_ground

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##                 Df Deviance    AIC
## + aircraft_num  1   78.164 106.41
## + height        1   95.059 123.30
## + pitch         1   97.006 125.25
## <none>              107.401 128.92
## + duration      1  107.296 135.54
## + no_pasg       1  107.375 135.62

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=104.84
## long_landing ~ speed_ground + aircraft_num

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##             Df Deviance     AIC
## + height     1   54.401  87.798
## <none>           78.164 104.836
## + pitch      1   75.176 108.572
## + duration   1   76.635 110.031
## + no_pasg    1   77.824 111.220

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Step:  AIC=83.94
## long_landing ~ speed_ground + aircraft_num + height

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##            Df Deviance    AIC
## <none>          54.401 83.942
## + pitch     1   51.580 87.844
## + duration  1   53.680 89.944
## + no_pasg   1   54.401 90.666
```

Step 8: Meeting with the FAA agent

- We will be modelling the variable landing distance using the two predictors - speed ground and aircraft numeric. They are the most important variables as they high association with our response variable.

- We observe that with a unit increase in speed ground, the odds ratio increases by 1.795 when the variable aircraft numeric is kept constant.

- We observe that with a unit increase in aircraft numeric (Basically here we are changing the aircraft type) the odds ratio increases by 0.039 when the variable speed_ground is kept constant.

```
presentation_model <- glm(long_landing ~ speed_ground + aircraft_num, family
= binomial, data = flights_1)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

odds_ratio_presentation <- c(
exp(summary(presentation_model)$coefficients[2,1]),
exp(summary(presentation_model)$coefficients[3,1]))

summary(presentation_model)

##
## Call:
```

```
## glm(formula = long_landing ~ speed_ground + aircraft_num, family =
binomial,
##     data = flights_1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.28368  -0.01417  -0.00039   0.00000   2.56541
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -57.53370    8.24419  -6.979 2.98e-12 ***
## speed_ground   0.58534    0.08441   6.934 4.08e-12 ***
## aircraft_num  -3.23679    0.71189  -4.547 5.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 626.946  on 831  degrees of freedom
## Residual deviance:  84.665  on 829  degrees of freedom
## AIC: 90.665
##
## Number of Fisher Scoring iterations: 10
```

Step 9 : Repeating Steps 1-7 for the binary variable Risky Landing

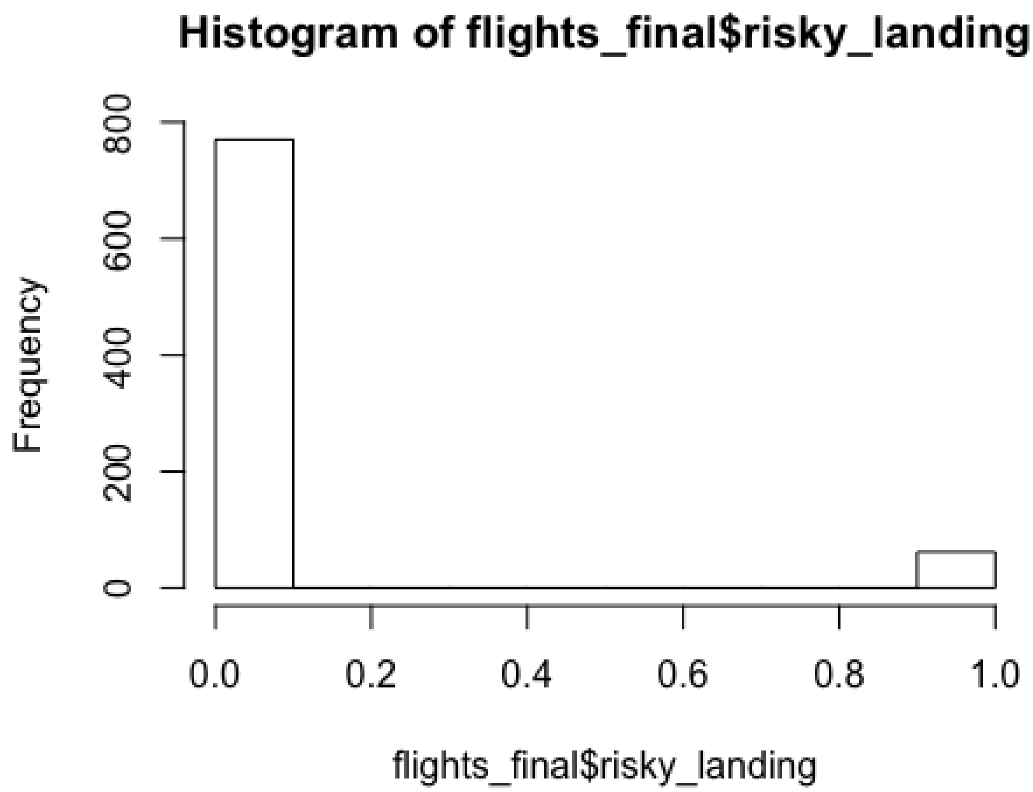
Step 1 (Risk Landing)

- A binary response of risky landing is created based on the varible distance. If the distance is greater than 3000 then variable risky landing will be 1 else it will be 0.

Step 2 (Risky Landing): Histogram showing the distribution of risky_landing

- It is observed that 62 observations of risky_landing have the value 1 and the rest 770 have the value 0.

```
hist(flights_final$risky_landing)
```



Step 3 (Risky Landing) : Fitting single-factor logistic regression

- The variable Risky Landing is logistically regressed with all the variables one by one.
- Later the results of all the regression models are tabulated in table 2 that contains the size regression coefficient, direction of coeficient, odds ratio and p values.
- Using p- values from the table 2, it is observed that the significant predictor variables are speed_air, speed_ground, and aircraft_num.

```
## Fitting single-factor logistic regression using each variable

duration_1 <-  glm(risky_landing ~ duration, family = binomial, data =
flights_final)
no_pasg_1 <- glm(risky_landing ~ no_pasg, family = binomial, data =
flights_final)
speed_ground_1 <- glm(risky_landing ~ speed_ground, family = binomial, data =
flights_final)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

speed_air_1 <- glm(risky_landing ~ speed_air, family = binomial, data =
flights_final)
height_1 <- glm(risky_landing ~ height, family = binomial, data =
```

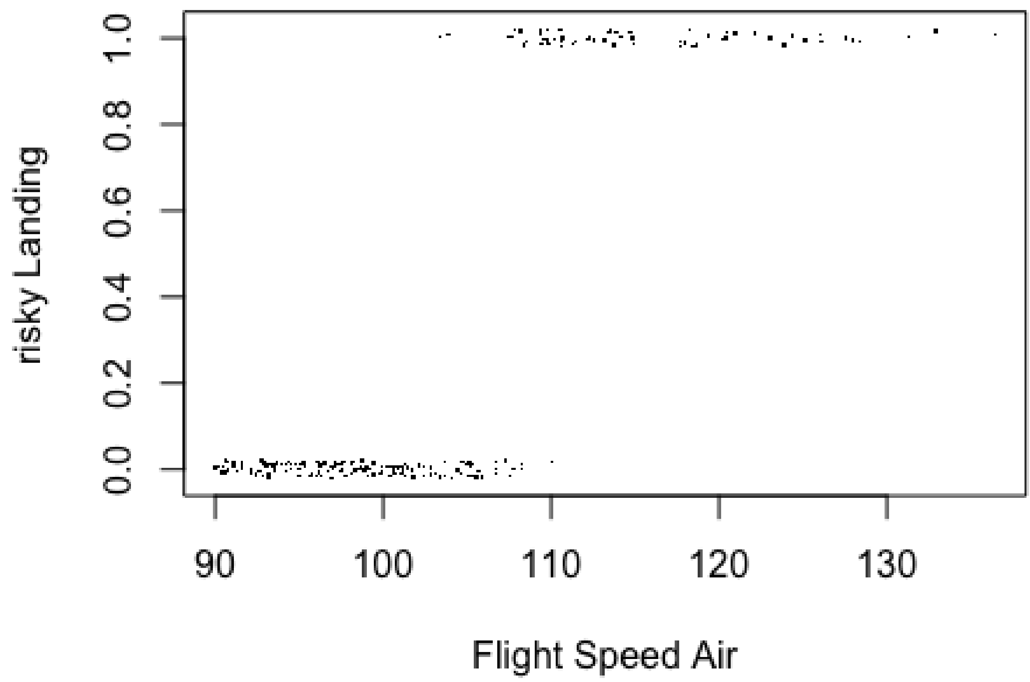
```
flights_final)
pitch_1 <- glm(risky_landing ~ pitch, family = binomial, data =
flights_final)
aircraft_num_1 <- glm(risky_landing ~ aircraft_num, family = binomial, data =
flights_final)


##Calculating odds ratio

odds_ratio_1 <- c(
exp(summary(duration_1)$coefficients[2,1]),
exp(summary(no_pasg_1)$coefficients[2,1]),
exp(summary(speed_ground_1)$coefficients[2,1]),
exp(summary(speed_air_1)$coefficients[2,1]),
exp(summary(height_1)$coefficients[2,1]),
exp(summary(pitch_1)$coefficients[2,1]),
exp(summary(aircraft_num_1)$coefficients[2,1]))


## P values

p_values_1 <- c(
summary(duration_1)$coefficients[2,4],
summary(no_pasg_1)$coefficients[2,4],
summary(speed_ground_1)$coefficients[2,4],
summary(speed_air_1)$coefficients[2,4],
summary(height_1)$coefficients[2,4],
summary(pitch_1)$coefficients[2,4],
summary(aircraft_num_1)$coefficients[2,4])

## Regression Coefficients

regression_coefficients_1 <- c(
summary(duration_1)$coefficients[2,1],
summary(no_pasg_1)$coefficients[2,1],
summary(speed_ground_1)$coefficients[2,1],
summary(speed_air_1)$coefficients[2,1],
summary(height_1)$coefficients[2,1],
summary(pitch_1)$coefficients[2,1],
summary(aircraft_num_1)$coefficients[2,1])

Table_2 <- data.frame(variable_names, regression_coefficients_1,
odds_ratio_1, coef_direction = ifelse(regression_coefficients_1 < 0,
"Negative", "Positive") , p_values_1)

Table_2

##   variable_names regression_coefficients_1 odds_ratio_1 coef_direction
## 1       duration             -0.0013826041    0.9986184       Negative
```

```
## 2        no_pasg              -0.0238804478    0.9764024       Negative
## 3   speed_ground               0.6142187540    1.8482121       Positive
## 4      speed_air               0.8704019017    2.3878703       Positive
## 5         height               0.0001493793    1.0001494       Positive
## 6          pitch               0.3755782194    1.4558330       Positive
## 7   aircraft_num              -1.0253058273    0.3586868       Negative
##     p_values_1
## 1 6.185950e-01
## 2 1.762083e-01
## 3 6.897975e-08
## 4 3.728025e-06
## 5 9.911654e-01
## 6 1.358414e-01
## 7 3.183632e-04
```

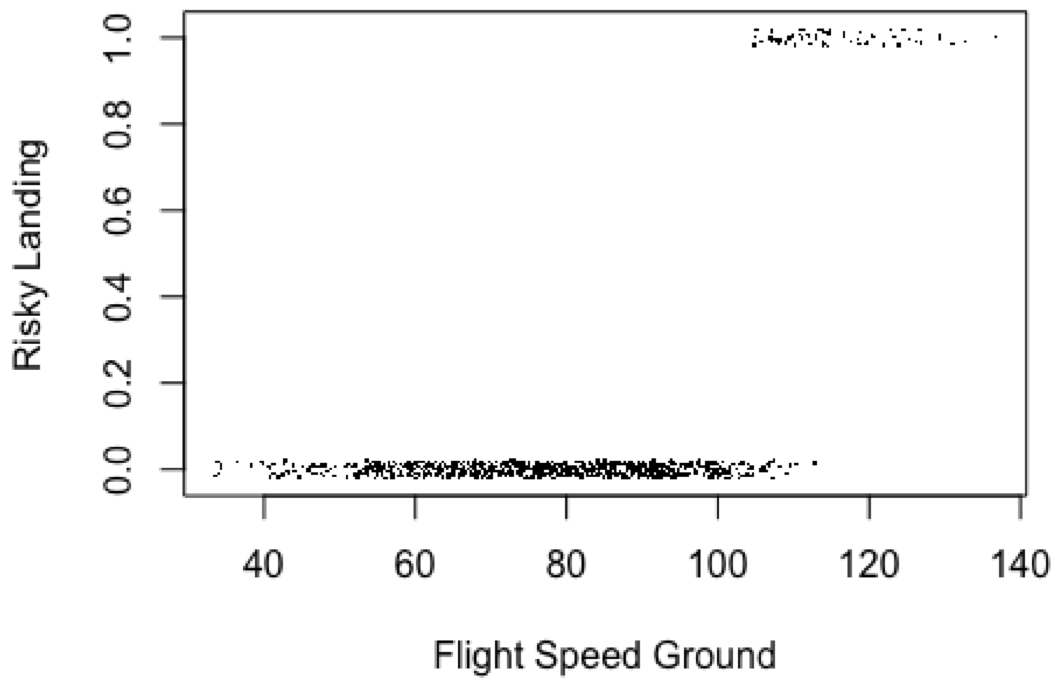
Step 4 (Risky Landing) : Seeing the association of Risky landing

- The significance of variables is checked using the p values. The models having p-values less than 0.05 are considered as significant.
- The significant predictor variables observed in table_1 are speed_air, speed_ground, and aircraft_num.

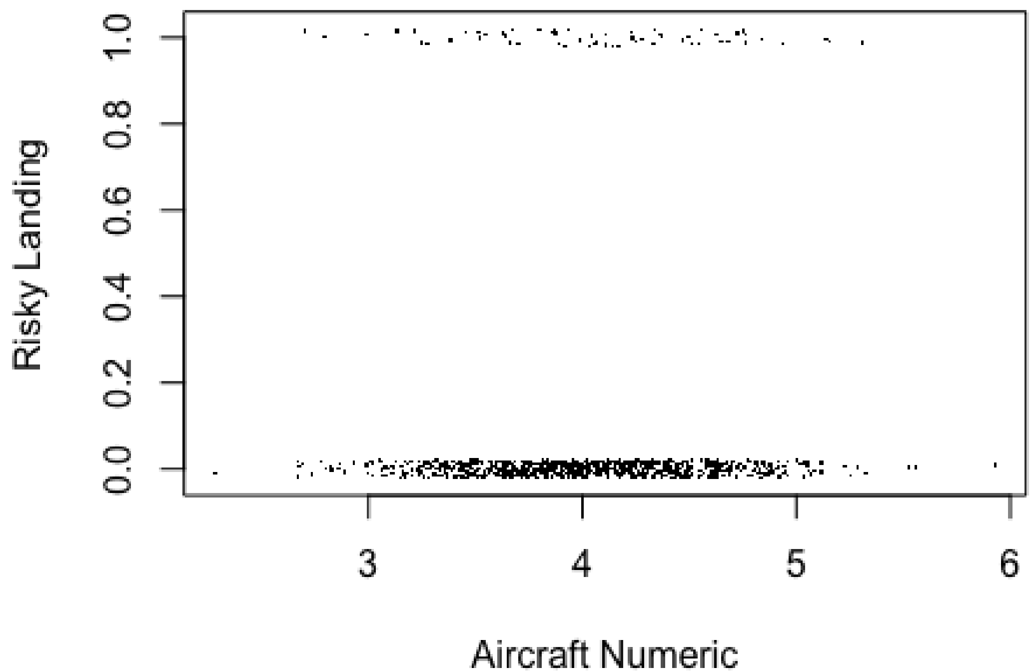
```
##Speed_air

plot(jitter(risky_landing,0.1) ~ jitter(speed_air), flights_final, xlab =
"Flight Speed Air", ylab = "risky Landing", pch = ".")
```




```
##Speed_ground

plot(jitter(risky_landing,0.1) ~ jitter(speed_ground), flights_final, xlab =
"Flight Speed Ground", ylab = "Risky Landing", pch = ".")
```




```
##Aircraft Numeric

plot(jitter(risky_landing,0.1) ~ jitter(pitch), flights_final, xlab =
"Aircraft Numeric", ylab = "Risky Landing", pch = ".")
```



Step 5 (Risky Landing) : Fitting the data with all variables together

- It ws observed in step 16 of Project part 1 that the speed air and speed ground were highly collinear. We used speed ground as predictor because the number of NA's in data were high for speed air. Also, speed ground was more significant than speeed air.

- We will now fit a logistic regression using two variables together. The varibles that we will use are speed_ground and aircraft numeric.

- The full model logistetic regression model tells us that wih a unit increase in Speed Ground the odds ratio will increase by 2.52 when all other variables are kept constant.

- The full model logistetic regression model tells us that wih a unit increase in Aircraft Numeric the odds ratio will increase by 0.017 when all other variables are kept constant.

```
full_model_1 <-  glm(risky_landing ~ speed_ground + aircraft_num, family =
binomial, data = flights_final)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Calculating odds ratio

odds_ratio_full_model_1 <- c(
```

```r
exp(summary(full_model_1)$coefficients[2,1]),
exp(summary(full_model_1)$coefficients[3,1]))

summary(full_model_1)

##
## Call:
## glm(formula = risky_landing ~ speed_ground + aircraft_num, family =
binomial,
##     data = flights_final)
##
## 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)  -98.0582    23.8303  -4.115 3.87e-05 ***
## speed_ground   0.9263     0.2248   4.121 3.78e-05 ***
## aircraft_num  -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: 441.251  on 831  degrees of freedom
## Residual deviance:  40.097  on 829  degrees of freedom
## AIC: 46.097
##
## Number of Fisher Scoring iterations: 12
```

Step 6 (Risky Landing) : Step Wise AIC

- We will use the Stepwise AIC funtion in R to do the variable selection for the full model of Logistic Regression.
- Before doing that we will remove the character variable aircraft type from the data frame as we have already coded it as binary. We will also remove the speed air variable as it has a lot of NULL values and it is highly collinear with speed ground.
- After applying the step AIC function to the model, is shows that it has lowest AIC of 45.71 when the variables speed_ground, aicraft_num and no_pasg are selected. The AIC for the model with variables speed_ground and aircraft_num is 46.1. Since this difference is small we will choose speed ground and aicraft_numeric as the predictor variables. Another reason behind that is we have already seen that no_pasg was less significant in the earlier steps.

```r
flights_2 <- dplyr::select(flights_final, duration, no_pasg, speed_ground,
height, pitch, aircraft_num, risky_landing)
```

```
GLM_long_landing_null_1 <- glm(risky_landing ~ 1, family = binomial, data =
flights_2)
GLM_long_landing_full_1 <- glm(risky_landing ~ ., family = binomial, data =
flights_2)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

fit1_GLM_1 <- step(GLM_long_landing_null_1, scope = list(lower
=GLM_long_landing_null_1,upper = GLM_long_landing_full_1),    direction =
'forward')

## Start:  AIC=443.25
## risky_landing ~ 1

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##                 Df Deviance    AIC
## + speed_ground  1    57.99  74.91
## + aircraft_num  1   416.49 433.41
## <none>              428.33 443.25
## + no_pasg       1   426.50 443.42
## + pitch         1   426.59 443.51
## + duration      1   428.08 445.00
## + height        1   428.32 445.24

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=62.93
## risky_landing ~ speed_ground

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
```

```
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##                Df Deviance    AIC
## + aircraft_num  1   39.955 46.898
## + pitch         1   51.634 58.576
## <none>              57.988 62.931
## + no_pasg       1   57.178 64.121
## + height        1   57.787 64.729
## + duration      1   57.951 64.893

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=46.1
## risky_landing ~ speed_ground + aircraft_num

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##            Df Deviance    AIC
## + no_pasg   1   37.559 45.700
## <none>          39.955 46.097
## + height    1   39.295 47.436
## + duration  1   39.757 47.898
## + pitch     1   39.783 47.924

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=45.71
## risky_landing ~ speed_ground + aircraft_num + no_pasg
```

```
## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##            Df Deviance    AIC
## <none>          37.559 45.707
## + height    1   36.985 47.133
## + pitch     1   37.304 47.452
## + duration  1   37.548 47.696
```

Step 7 (Risky Landing) : Step Wise BIC

- The step function in R can also be used with BIC as our parameter. We will give an extra argument 'k = log(nrow(flights_1))' in the step function. The use of this function for BIC was found through google search. Here is its link
- We observe similar kind of results in stepwise BIC as well. The minimum BIC of 60.27 is observed when the variable speed_ground and aircraft numerice are used as predictor variables.
- Therefore, the final variables that we will be using as predictors are speed ground and aircraft_numeric.

```
fit2_GLM_1 <- step(GLM_long_landing_null_1, scope = list(lower
=GLM_long_landing_null_1,upper = GLM_long_landing_full_1),    direction =
'forward', k = log(nrow(flights_1)))

## Start:  AIC=447.97
## risky_landing ~ 1

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                Df Deviance    AIC
## + speed_ground  1    57.99  84.35
## + aircraft_num  1   416.49 442.86
## <none>              428.33 447.97
## + no_pasg       1   426.50 452.87
```

```
## + pitch         1   426.59 452.96
## + duration      1   428.08 454.45
## + height        1   428.32 454.68

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=72.38
## risky_landing ~ speed_ground

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##                  Df Deviance    AIC
## + aircraft_num  1   39.955 61.069
## <none>              57.988 72.378
## + pitch         1   51.634 72.748
## + no_pasg       1   57.178 78.292
## + height        1   57.787 78.901
## + duration      1   57.951 79.065

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=60.27
## risky_landing ~ speed_ground + aircraft_num

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## using the 782/832 rows from a combined fit

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning in add1.glm(fit, scope$add, scale = scale, trace = trace, k = k, :
## glm.fit: fitted probabilities numerically 0 or 1 occurred

##             Df Deviance    AIC
## <none>           39.955 60.268
## + no_pasg    1   37.559 64.596
## + height     1   39.295 66.332
## + duration   1   39.757 66.794
## + pitch      1   39.783 66.820
```

Step 10 : Meeting the FAA agent

- We will be modelling the variable risky landing distance using the predictors - speed ground and aircraft numeirc. They are the most important variable as they have high association with our response variable.
- We observe that with a unit increase in speed ground, the odds ratio increases by 2.52 when other variables are kept constant.
- We observe that with a unit increase in aircraft_num, the odds ratio increases by 0.0179 when other variables are kept constant.

```
presentation_model_1 <- glm(risky_landing ~ speed_ground  + aircraft_num,
family = binomial, data = flights_2)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

odds_ratio_presentation_1 <- c(
exp(summary(presentation_model_1)$coefficients[2,1]),
exp(summary(presentation_model_1)$coefficients[3,1]))

summary(presentation_model_1)

##
## Call:
## glm(formula = risky_landing ~ speed_ground + aircraft_num, family =
binomial,
##     data = flights_2)
##
## 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)  -98.0582    23.8303  -4.115 3.87e-05 ***
## speed_ground   0.9263     0.2248   4.121 3.78e-05 ***
## aircraft_num  -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: 441.251  on 831  degrees of freedom
## Residual deviance:  40.097  on 829  degrees of freedom
## AIC: 46.097
##
## Number of Fisher Scoring iterations: 12
```

Step 11 : Comparison of Two Models

- For the prediction of probability of long landing we have used the variables speed of ground and the aircraft type. We observe that with a unit increase in speed ground, the odds ratio increases by 1.79 when the variable aircraft numeric is kept constant. We observe that with a unit increase in aircraft numeric (Basically here we are changing the aircraft type) the odds ratio increases by 0.039 when the variable speed_ground is kept constant.

- For the prediction of probability of long landing we will be using the variables speed of ground and aircraft numeric. We observe that with a unit increase in speed ground, the odds ratio increases by 2.52 when aircraft numeric is kept constant. While there is an increase in odds ratio by 0.0179 when aircraft_num is icreased by 1 unit keeping speed_ground constant.

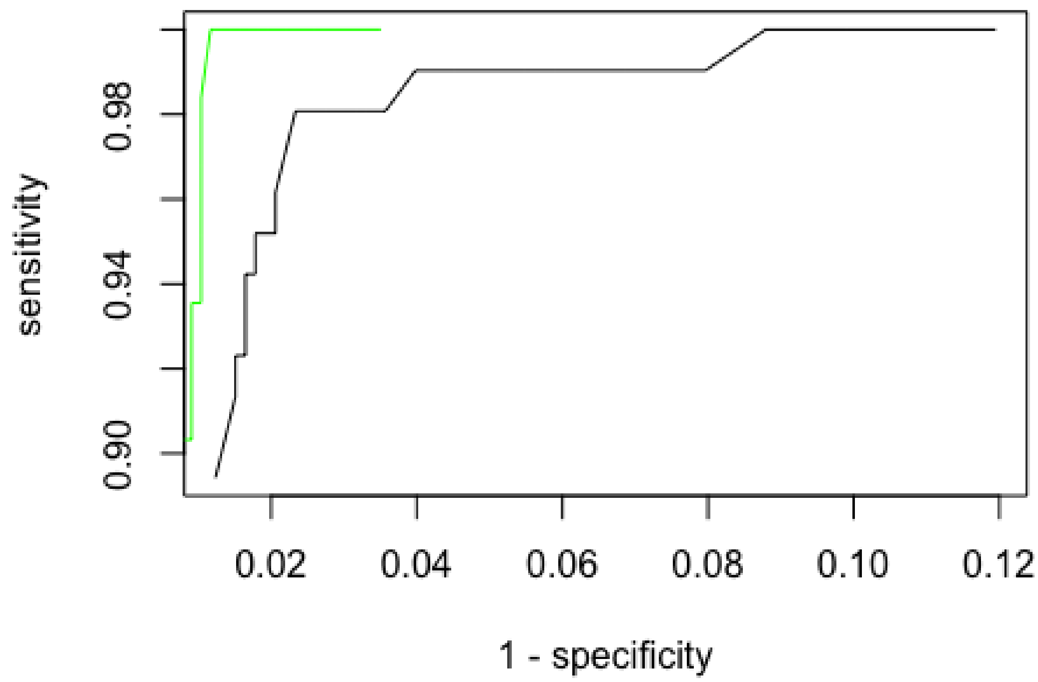
- Speed Air could have been a good predictor for both the binary variables as it also had a great assiciation with them. Owing to high number of null values we are unable to use that in our models.

Step 12 : ROC Curves

- After plotting the ROC Curves for our final models we observe that the model built for risky landing is better than the model built for long landing. The area under the curve for the former model is greater than the latter one.

```r
## Long Landing Model


thresh <- seq(0.01, 0.5, 0.01)

pred_prob <- predict(presentation_model, type = "response")
pred_prob_1 <- predict(presentation_model_1, type = "response")
```

```r
## Data Frames for Graphs

flights_1a <- data.frame(flights_1, pred_prob)

sensitivity <- specificity <- rep(NA, length(thresh))
for (j in seq(along = thresh)){
  pp <- ifelse(flights_1a$pred_prob < thresh[j], "no","yes")
  xx <- xtabs(~long_landing + pp, flights_1a)
  specificity[j] <- xx[1,1]/(xx[1,1] + xx[1,2])
  sensitivity[j] <- xx[2,2]/(xx[2,1] + xx[2,2])
}

flights_2a <- data.frame(flights_2, pred_prob_1)

sensitivity_1 <- specificity_1 <- rep(NA, length(thresh))
for (j in seq(along = thresh)){
  pp_1 <- ifelse(flights_2a$pred_prob_1 < thresh[j], "no","yes")
  xx_1 <- xtabs(~risky_landing + pp_1, flights_2a)
  specificity_1[j] <- xx_1[1,1]/(xx_1[1,1] + xx_1[1,2])
  sensitivity_1[j] <- xx_1[2,2]/(xx_1[2,1] + xx_1[2,2])
}


plot(1-specificity, sensitivity, type = "l"); abline(0,1, lty = 2)
lines(1-specificity_1,sensitivity_1,col="green")
```



Step 13 : Predicting Probability for given observation

- We will now predict the probabilty and confidence intervals for the given observation and for both the variables long_landing and risky_landing.
- Since the aircraft type in the given observation is 'Boeing', the aircraft numeric will be equal to 1.
- The Probability of long landing for this observation predicted by the model is 99.99%. The confidence interval is in between 0.9998 and 1.0001.
- The Probability of long landing for this observation predicted by the model is 99.97%. The confidence interval is in between 0.9989 and 1.0007

```
new_obs <- data.frame(speed_ground = 115, aircraft_num = 0)

## Prediction of Probability of Long Landing
predict(presentation_model, newdata = new_obs, type = "response", se = T)

## $fit
##         1
## 0.9999434
##
## $se.fit
```

```
##            1
## 8.630534e-05
##
## $residual.scale
## [1] 1
```

```r
## Confidence Interval of Long Landing

round(c(0.9999434 - 1.96*8.630534e-05, 0.9999434 + 1.96*8.630534e-05 ), 4)
```

```
## [1] 0.9998 1.0001
```

```r
## Prediction of probability of Risky Landing

predict(presentation_model_1, newdata = new_obs, type = "response", se = T)
```

```
## $fit
##        1
## 0.999789
##
## $se.fit
##            1
## 0.0004408113
##
## $residual.scale
## [1] 1
```

```r
## Confidence Interval of Long Landing

round(c(0.999789  - 1.96*0.0004408113, 0.999789  + 1.96*0.0004408113), 4)
```

```
## [1] 0.9989 1.0007
```

Step 14 : Fitting the Probit and Hazard model for the variable Risky Landing

- We will be using the same variables i.e. speed_ground and aircraft numeric as predictor variables which were found as important in steps 9 and 10.
- After fitting the models, we see that the size coefficients of the earlier model is almost twice when compared with probit and complementary log log model.
- The sizes of coeffients is almost the same for probit model and the complementary log log model.

```r
## Fitting a Probit Model

presentation_model_1_probit <- glm(risky_landing ~ speed_ground +
aircraft_num, family = binomial(link = probit), flights_2)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

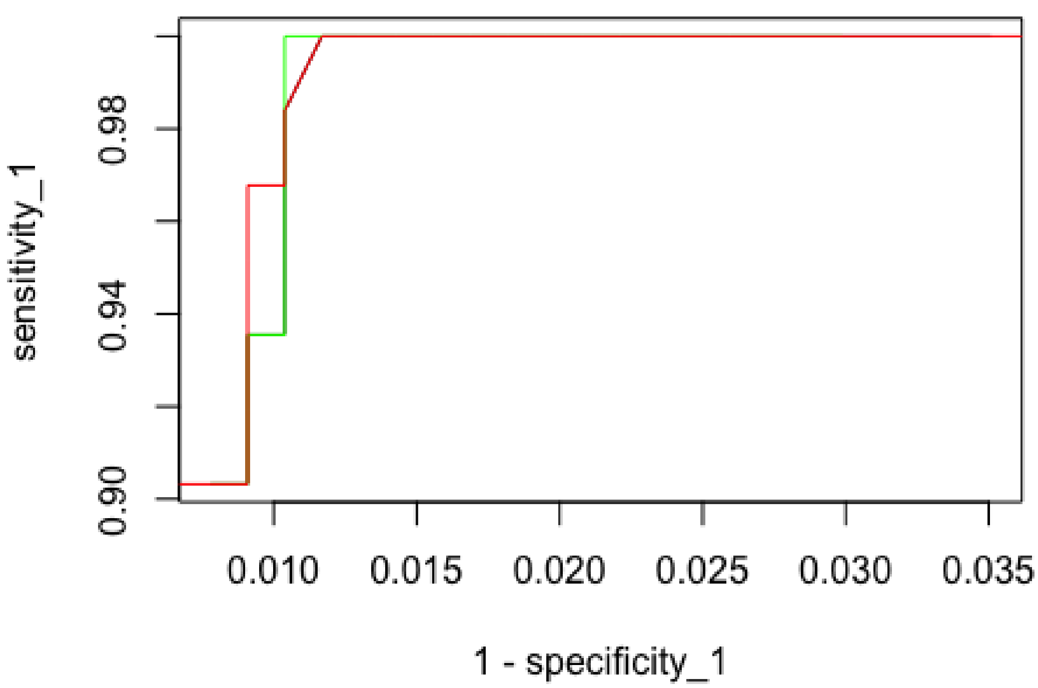
```r
## Fitting a C.Log Log Model

presentation_model_1_cloglog <- glm(risky_landing ~ speed_ground +
aircraft_num, family = binomial(link = cloglog), flights_2)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Comparing the models of risky landing

round(coef(presentation_model_1), 3)

##  (Intercept) speed_ground aircraft_num
##      -98.058        0.926       -4.019

round(coef(presentation_model_1_probit), 3)

##  (Intercept) speed_ground aircraft_num
##      -56.336        0.532       -2.357

round(coef(presentation_model_1_cloglog), 3)

##  (Intercept) speed_ground aircraft_num
##      -66.367        0.622       -2.898
```

Step 15 : Comparing the ROC curves for all the three models

- After comparing the graphs of all three models we observe that the highest AUC is for the complementary log log model, then the probit model followed by the general linear model.
- The green graph represents probit model, the red represents the complementary log log model.

```r
pred_prob_1_probit <- predict(presentation_model_1_probit, type = "response")
pred_prob_1_cloglog <- predict(presentation_model_1_cloglog, type =
"response")



flights_3a <- data.frame(flights_2, pred_prob_1,
pred_prob_1_probit,pred_prob_1_cloglog)

sensitivity_1_probit <- specificity_1_probit <- rep(NA, length(thresh))
for (j in seq(along = thresh)){
  pp_1_probit <- ifelse(flights_3a$pred_prob_1_probit < thresh[j],
"no","yes")
  xx_1_probit <- xtabs(~risky_landing + pp_1_probit, flights_3a)
  specificity_1_probit[j] <- xx_1_probit[1,1]/(xx_1_probit[1,1] +
xx_1_probit[1,2])
  sensitivity_1_probit[j] <- xx_1_probit[2,2]/(xx_1_probit[2,1] +
xx_1_probit[2,2])
}
```

```
sensitivity_1_cloglog <- specificity_1_cloglog <- rep(NA, length(thresh))
for (j in seq(along = thresh)){
  pp_1_cloglog <- ifelse(flights_3a$pred_prob_1_cloglog < thresh[j],
"no","yes")
  xx_1_cloglog <- xtabs(~risky_landing + pp_1_cloglog, flights_3a)
  specificity_1_cloglog[j] <- xx_1_cloglog[1,1]/(xx_1_cloglog[1,1] +
xx_1_cloglog[1,2])
  sensitivity_1_cloglog[j] <- xx_1_cloglog[2,2]/(xx_1_cloglog[2,1] +
xx_1_cloglog[2,2])
}


plot(1-specificity_1, sensitivity_1, type = "l"); abline(0,1, lty = 2)
lines(1-specificity_1_probit,sensitivity_1_probit,col="green")
lines(1-specificity_1_cloglog,sensitivity_1_cloglog,col="red")
```



Step 16: Top 5 Risky Landings

- We will be using the 'top_n' function in R to do this. Tis was figured out by google search. Here is its link
- All the 3 models point towards same set of 5 flights when sorted by the higest probabilities.

```r
## Top 5 Flights - General Linear model
top_n(flights_3a, 5, pred_prob_1)
```

```
##    duration no_pasg speed_ground   height    pitch aircraft_num
## 1 161.89247      72     129.2649 33.94900 4.139951            0
## 2 119.92455      64     136.6592 44.28611 4.169404            0
## 3 154.52460      67     129.3072 23.97850 5.154699            0
## 4  63.32952      52     132.7847 18.17703 4.110664            0
## 5 153.83445      61     126.8393 20.54783 4.334558            0
##   risky_landing pred_prob_1 pred_prob_1_probit pred_prob_1_cloglog
## 1             1           1                  1                   1
## 2             1           1                  1                   1
## 3             1           1                  1                   1
## 4             1           1                  1                   1
## 5             1           1                  1                   1
```

```r
## Top 5 Flights - Probit model
top_n(flights_3a, 5, pred_prob_1_probit)
```

```
##     duration no_pasg speed_ground   height    pitch aircraft_num
## 1  116.98454      67     122.7566 30.21657 3.213703            0
## 2  161.89247      72     129.2649 33.94900 4.139951            0
## 3  209.10635      58     124.5699 40.10112 4.648428            0
## 4  119.92455      64     136.6592 44.28611 4.169404            0
## 5  197.54635      68     126.6692 23.76423 2.993151            0
## 6  232.79386      56     123.9569 26.36755 4.061951            0
## 7  154.52460      67     129.3072 23.97850 5.154699            0
## 8   63.32952      52     132.7847 18.17703 4.110664            0
## 9   99.68150      61     121.8371 33.18460 3.867476            0
## 10 153.83445      61     126.8393 20.54783 4.334558            0
## 11 131.73110      60     131.0352 28.27797 3.660194            1
## 12 137.58573      66     126.2443 35.17570 2.701924            1
##    risky_landing pred_prob_1 pred_prob_1_probit pred_prob_1_cloglog
## 1              1   0.9999998                  1                   1
## 2              1   1.0000000                  1                   1
## 3              1   1.0000000                  1                   1
## 4              1   1.0000000                  1                   1
## 5              1   1.0000000                  1                   1
## 6              1   0.9999999                  1                   1
## 7              1   1.0000000                  1                   1
## 8              1   1.0000000                  1                   1
## 9              1   0.9999996                  1                   1
## 10             1   1.0000000                  1                   1
## 11             1   1.0000000                  1                   1
## 12             1   0.9999996                  1                   1
```

```r
## Top 5 Flights - complementary log-log model
top_n(flights_3a, 5, pred_prob_1_cloglog)
```

```
##     duration no_pasg speed_ground   height    pitch aircraft_num
## 1  163.90650      55     119.3805 38.55854 3.701449            0
```

```
## 2  140.23631      65     118.7420 19.85619 4.646266            0
## 3  130.46356      52     116.7134 36.19553 3.894352            0
## 4  116.98454      67     122.7566 30.21657 3.213703            0
## 5  161.89247      72     129.2649 33.94900 4.139951            0
## 6  205.87361      62     113.9963 34.44342 3.873845            0
## 7  209.10635      58     124.5699 40.10112 4.648428            0
## 8  127.99133      59     114.2927 25.46814 5.138243            0
## 9  113.36296      56     113.9640 44.73546 3.937906            0
## 10 119.92455      64     136.6592 44.28611 4.169404            0
## 11 197.17730      58     113.8891 33.45538 4.233058            0
## 12 197.54635      68     126.6692 23.76423 2.993151            0
## 13 232.79386      56     123.9569 26.36755 4.061951            0
## 14 272.03906      59     118.9227 15.04935 4.106572            0
## 15 277.17601      52     119.6539 25.18276 4.934241            0
## 16 164.23895      59     113.0295 38.34827 3.276835            0
## 17 124.48006      60     114.4807 45.07767 4.334137            0
## 18 109.45172      66     117.6406 35.91004 4.058218            0
## 19 154.52460      67     129.3072 23.97850 5.154699            0
## 20 166.10453      48     116.5925 13.26324 3.133959            0
## 21  99.19386      60     119.6775 27.55802 3.640565            0
## 22  63.32952      52     132.7847 18.17703 4.110664            0
## 23  99.68150      61     121.8371 33.18460 3.867476            0
## 24 153.83445      61     126.8393 20.54783 4.334558            0
## 25 131.73110      60     131.0352 28.27797 3.660194            1
## 26 158.53503      62     118.5190 25.78507 3.523655            1
## 27 140.67120      48     120.4548 30.35151 4.371072            1
## 28 137.58573      66     126.2443 35.17570 2.701924            1
## 29 140.45311      75     120.4189 31.26345 2.796731            1
## 30 175.51443      49     125.2123 22.52478 4.365772            1
## 31 220.05713      61     120.5579 15.66566 4.111265            1
## 32  98.50031      66     123.3105 22.32718 4.276710            1
##    risky_landing pred_prob_1 pred_prob_1_probit pred_prob_1_cloglog
## 1              1   0.9999964          1.0000000                   1
## 2              1   0.9999934          1.0000000                   1
## 3              1   0.9999568          1.0000000                   1
## 4              1   0.9999998          1.0000000                   1
## 5              1   1.0000000          1.0000000                   1
## 6              1   0.9994655          0.9999928                   1
## 7              1   1.0000000          1.0000000                   1
## 8              1   0.9995938          0.9999965                   1
## 9              1   0.9994493          0.9999922                   1
## 10             1   1.0000000          1.0000000                   1
## 11             1   0.9994097          0.9999907                   1
## 12             1   1.0000000          1.0000000                   1
## 13             1   0.9999999          1.0000000                   1
## 14             1   0.9999944          1.0000000                   1
## 15             1   0.9999972          1.0000000                   1
## 16             1   0.9986922          0.9999342                   1
## 17             1   0.9996587          0.9999978                   1
## 18             1   0.9999817          1.0000000                   1
```

```
## 19              1   1.0000000         1.0000000                     1
## 20              1   0.9999517         1.0000000                     1
## 21              1   0.9999972         1.0000000                     1
## 22              1   1.0000000         1.0000000                     1
## 23              1   0.9999996         1.0000000                     1
## 24              1   1.0000000         1.0000000                     1
## 25              1   1.0000000         1.0000000                     1
## 26              1   0.9995491         0.9999943                     1
## 27              1   0.9999249         1.0000000                     1
## 28              1   0.9999996         1.0000000                     1
## 29              1   0.9999224         1.0000000                     1
## 30              1   0.9999991         1.0000000                     1
## 31              1   0.9999318         1.0000000                     1
## 32              1   0.9999947         1.0000000                     1
```

Step 17: Prediction of probability and its confidence intervals for the observation in step 13

- The Probability of risky landing for the observation in step 13 predicted by the probit model is 99.99%. The confidence interval is in between 0.9998 and 1.0001.
- The Probability of long landing for this observation predicted by the model is 99.97%. The confidence interval is in between 0.99999 and 1.00001.
- The Probability of long landing for this observation predicted by the model is 100%. The confidence interval is in between 0.99999 and 1.00005.

```
## Prediction of probability of Risky Landing using probit model

predict(presentation_model_1_probit, newdata = new_obs, type = "response", se
= T)

## $fit
##         1
## 0.9999994
##
## $se.fit
##            1
## 3.153557e-06
##
## $residual.scale
## [1] 1

## Confidence Interval of Risky Landing using probit model

round(c(0.9999994   - 1.96*3.153557e-06, 0.9999994   + 1.96*3.153557e-06), 5)

## [1] 0.99999 1.00001

## Prediction of probability of Risky Landing using Complementary Log Log
model
```

```r
predict(presentation_model_1_cloglog, newdata = new_obs, type = "response",
se = T)

## $fit
## 1
## 1
##
## $se.fit
##            1
## 2.605523e-16
##
## $residual.scale
## [1] 1

## Confidence Interval of Risky Landing using Complementary Log Log model

round(c(1   - 1.96*2.605523e-16, 1   + 1.96*2.605523e-16), 6)

## [1] 1 1
```