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.

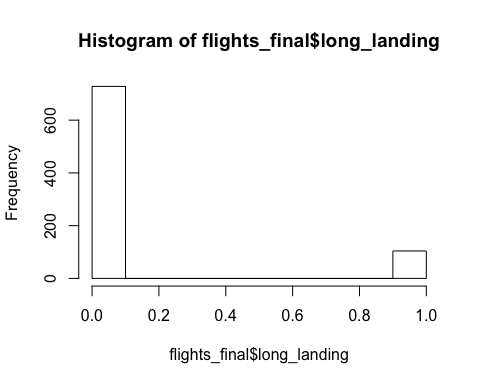
## Adding Binary Variables   
  
flights\_final$long\_landing <- ifelse(flights\_final$distance > 2500, 1, 0)  
  
flights\_final$risky\_landing <- ifelse(flights\_final$distance > 3000, 1, 0)  
  
## Discarding the Continous variable 'distance'  
  
flights\_final$distance <- NULL

# 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

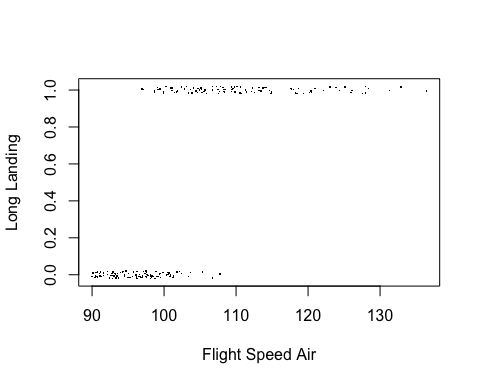
speed\_air <- glm(long\_landing ~ speed\_air, family = binomial, data = flights\_final)  
height <- glm(long\_landing ~ height, family = binomial, data = flights\_final)  
pitch <- glm(long\_landing ~ pitch, family = binomial, data = flights\_final)  
aircraft\_num <- glm(long\_landing ~ aircraft\_num, 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

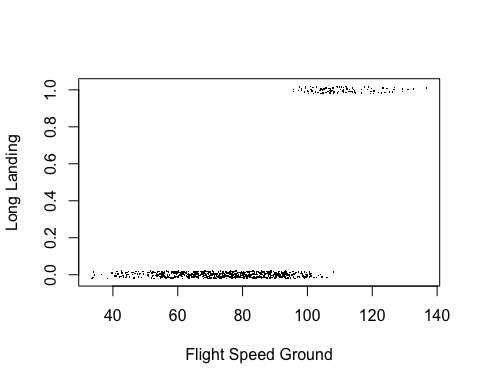
## Step 4 : Seeing the association of long landing

* 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.

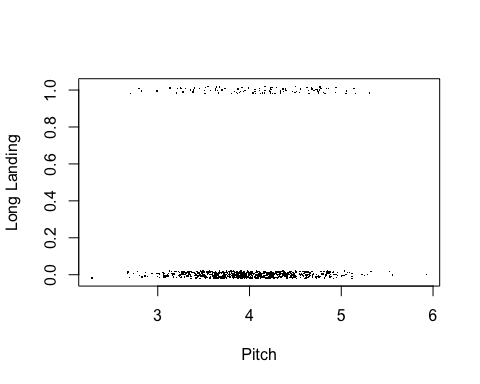
##Speed\_air  
  
plot(jitter(long\_landing,0.1) ~ jitter(speed\_air), flights\_final, xlab = "Flight Speed Air", ylab = "Long Landing", pch = ".")



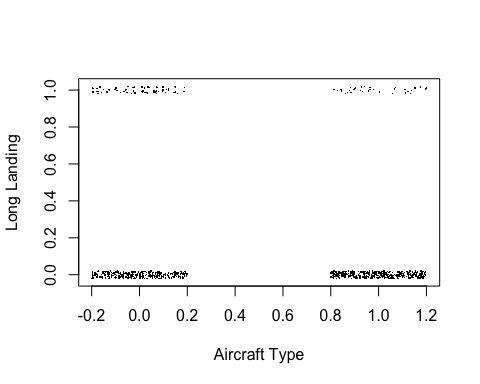
##Speed\_ground  
  
plot(jitter(long\_landing,0.1) ~ jitter(speed\_ground), flights\_final, xlab = "Flight Speed Ground", ylab = "Long Landing", pch = ".")



##Pitch  
  
plot(jitter(long\_landing,0.1) ~ jitter(pitch), flights\_final, xlab = "Pitch", ylab = "Long Landing", pch = ".")



##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

## 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.

## 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](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/step)
* 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

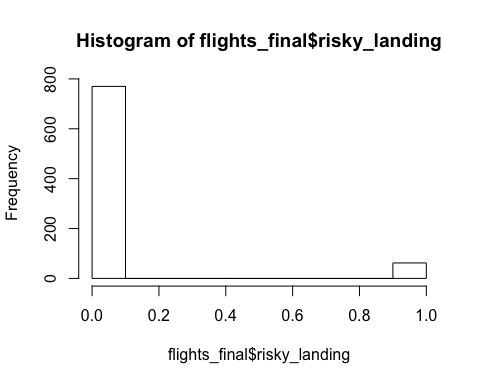
## Step 1 (Risk Landing)

* 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

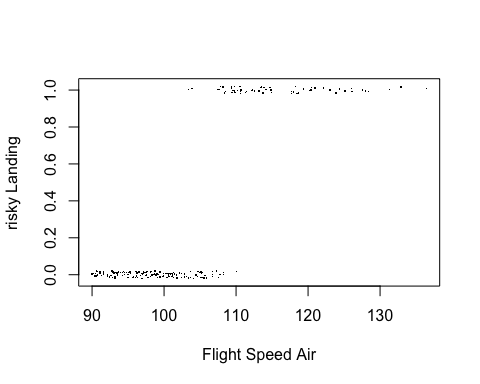
speed\_air\_1 <- glm(risky\_landing ~ speed\_air, family = binomial, data = flights\_final)  
height\_1 <- glm(risky\_landing ~ height, family = binomial, data = flights\_final)  
pitch\_1 <- glm(risky\_landing ~ pitch, family = binomial, data = flights\_final)  
aircraft\_num\_1 <- glm(risky\_landing ~ 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

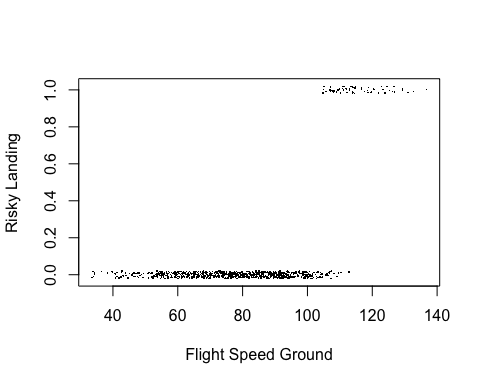
## Step 4 (Risky Landing) : Seeing the association of Risky landing

* 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.

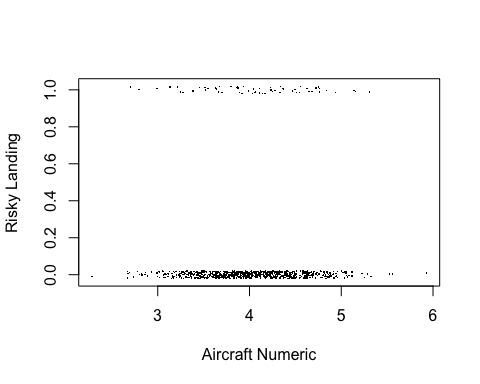
##Speed\_air  
  
plot(jitter(risky\_landing,0.1) ~ jitter(speed\_air), flights\_final, xlab = "Flight Speed Air", ylab = "risky Landing", pch = ".")



##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(  
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.

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](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/step)
* 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

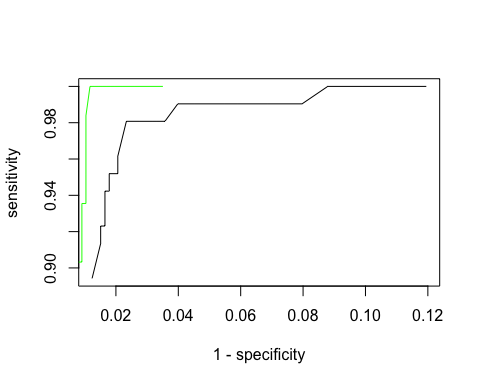
## Step 11 : Comparison of Two Models

* 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.

## 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")  
  
  
## 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

## 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

## 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

## 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.

## 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

## 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

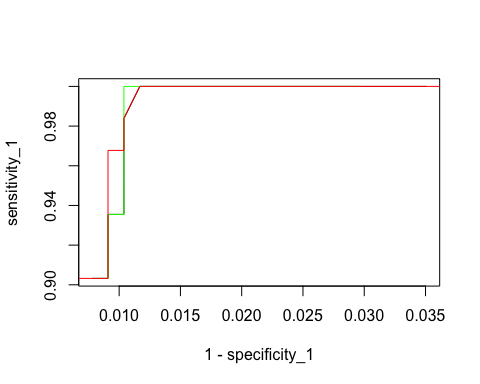
round(coef(presentation\_model\_1\_cloglog), 3)

## (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.

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](https://dplyr.tidyverse.org/reference/top_n.html)
* All the 3 models point towards same set of 5 flights when sorted by the higest probabilities.

## 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

## 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

## 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   
  
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