BANA 7042 STATISTICAL MODELING - PROJECT PART - 3

Name: Syed Imad Husain UCID: M12958531

Answer the following questions with support of analytic evidence.
Please provide

- The R code (how do you realize it?)
- The R output (Copy and paste only those relevant)
- Your observations (What do you observe from the output?)
- Your conclusion/decision/action

Contents

Question 1. Modeling Multinomial Data

Again, please work on the cleaned FAA data set you prepared by carrying out Steps 1-9 in Part 1 of the project. Create a multinomial variable and attach it to your data set.

- Y = 1 if distance < 1000
- Y = 2 if 1000 < = distance < 2500
- Y = 3 otherwise

Discard the continuous data for “distance”, and assume we are given this multinomial response only. In your meeting with an FAA agent who wants to know “what are risk factors in the landing process and how do they influence its occurrence?”, you are allowed to present:

- One model
- One table
- No more than five figures
- No more than five bullet statements
- Please use statements that she can understand

What model/table/figures/statements would you include in your presentation? Be selective!

Step 1. Cleaning Dataset & Discretizing Response Variable

Code

```
# step 0 create cleaned data set
#import data
library(readxl)
library(tidyverse)
library(magrittr)
FAA1 <- read_excel("FAA1.xls")
FAA2 <- read_excel("FAA2.xls")
#remove duplicates if any within data set
FAA1 <- unique(FAA1)
FAA2 <- unique(FAA2)
#finally create unique dataset
merged <- rbind(FAA1[,-2],FAA2)
merged <- unique(merged)
FAA <- merge(merged,FAA1,by=names(merged), all.x=TRUE)
```

```
#removing abnormal values
faa_clean <- FAA %>% select(names(FAA)) %>%
  filter(replace(duration,is.na(duration),60) > 40 &
           (speed_ground > 30 &
              speed_ground < 140) &
           (replace(speed_air,is.na(speed_air),60) > 30 &
              replace(speed_air,is.na(speed_air),60) < 140) &
           height >= 6 &
           distance < 6000   )
#faa is the final dataset to work on
faa <- faa_clean
#Discretization of Landing Distance
faa$Y <- 3
faa[which(faa$distance < 1000),"Y"] <- 1
faa[which(faa$distance < 2500 & faa$distance >= 1000),"Y"] <- 2
#Custom Encoded of Variables
#faa$Y <- ordered(factor(faa$Y, levels = c(1,2,3)))
faa$aircraft <- as.factor(faa$aircraft)
#Droping actual Landing Distance variable
table(faa$Y)
faa <- faa[,-7]
#Distribution of Y
library(ggplot2)
ggplot(data = faa, aes(x = Y,fill=Y)) +
      geom_bar(stat ="count")
names(faa)
```

Output

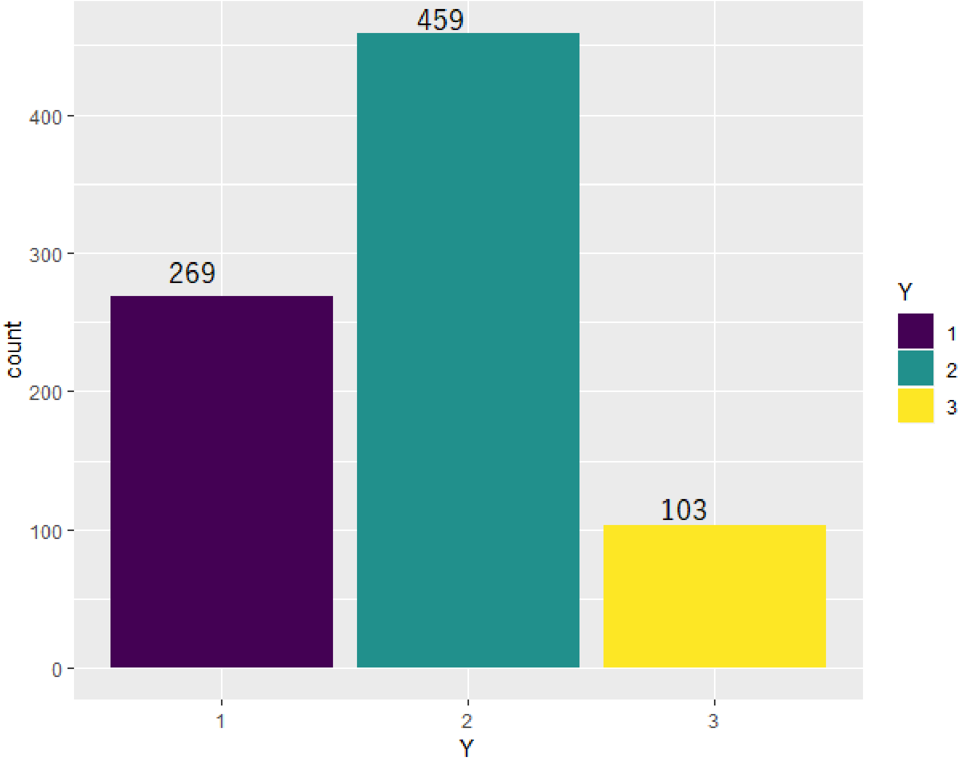


Table & Figure: 1 Distribution of multinomial response variable

Observation

The final cleaned data set has 8 variables and 831 observations. Y = 2, i.e. the second category has the greatest number of observations

Conclusion

The dataset is now ready for performing

Step 2. Statistical Significance of Individual variables

Code

```r
#building individual models to
aircraft.model<- multinom(Y ~ aircraft,data = faa)
no_pasg.model<- multinom(Y ~ no_pasg,data = faa)
speed_ground.model<- multinom(Y ~ speed_ground,data = faa)
speed_air.model<- multinom(Y ~ speed_air,data = faa)
height.model<- multinom(Y ~ height      ,data = faa)
pitch.model<- multinom(Y ~ pitch,data = faa)
duration.model<- multinom(Y ~ duration,data = faa)

z.aircraft<-
summary(aircraft.model)$coefficients/summary(aircraft.model)$standard.errors
z.no_pasg<-
summary(no_pasg.model)$coefficients/summary(no_pasg.model)$standard.errors
z.speed_ground<-
summary(speed_ground.model)$coefficients/summary(speed_ground.model)$standard.errors
z.speed_air<-
summary(speed_air.model)$coefficients/summary(speed_air.model)$standard.errors
z.height<- summary(height.model)$coefficients/summary(height.model)$standard.errors
z.pitch<- summary(pitch.model)$coefficients/summary(pitch.model)$standard.errors
z.duration<-
summary(duration.model)$coefficients/summary(duration.model)$standard.errors
z <- summary(aircraft.model)$coefficients/summary(aircraft.model)$standard.errors

# 2-tailed Wald z tests to test significance of coefficients
p.aircraft <- (1 - pnorm(abs(z.aircraft), 0, 1)) * 2
p.no_pasg <- (1 - pnorm(abs(z.no_pasg), 0, 1)) * 2
p.speed_ground <- (1 - pnorm(abs(z.speed_ground), 0, 1)) * 2
p.speed_air <- (1 - pnorm(abs(z.speed_air), 0, 1)) * 2
p.height <- (1 - pnorm(abs(z.height), 0, 1)) * 2
p.pitch <- (1 - pnorm(abs(z.pitch), 0, 1)) * 2
p.duration <- (1 - pnorm(abs(z.duration), 0, 1)) * 2

sum(p.aircraft[,2]>0.05) # significant
sum(p.no_pasg[,2]>0.05) # not significant
sum(p.speed_ground[,2]>0.05) #significant
sum(p.speed_air[2]>0.05) # significant
sum(p.height[,2]>0.05) #significant
sum(p.pitch[,2]>0.05) #not significant overall
sum(p.duration[,2]>0.05) #not significant overall
#visualize correlation with important factors
library(ggpubr)
g_ground <- ggplot(data <- faa,aes(x=speed_ground,fill=factor(Y)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
               colour=factor(Y)),alpha = 0.5)
```

```r
g_height <- ggplot(data <- faa,aes(x=height,fill=factor(Y)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
              colour=factor(Y)),alpha = 0.5)

beoing.index <- which(faa$aircraft=='boeing')

g_aircraft.b <- ggplot(data <- faa[beoing.index,],
                     aes(x=Y,fill=factor(Y)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
               colour=factor(Y)),alpha = 0.5)

g_aircraft.a <- ggplot(data <- faa[-beoing.index,],
                       aes(x=Y,fill=factor(Y)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                  colour=factor(Y)),alpha = 0.5)

ggarrange(g_ground,g_height,g_aircraft.b,g_aircraft.a, ncol = 2, nrow = 2)
```

Observation

Variable	Significance
aircraft	significant
no_pasg	not significant
speed_ground	significant
speed_air	significant
height	significant
pitch	not significant overall
duration	not significant overall

Table & Figure: 2 Individual Statistical Significance

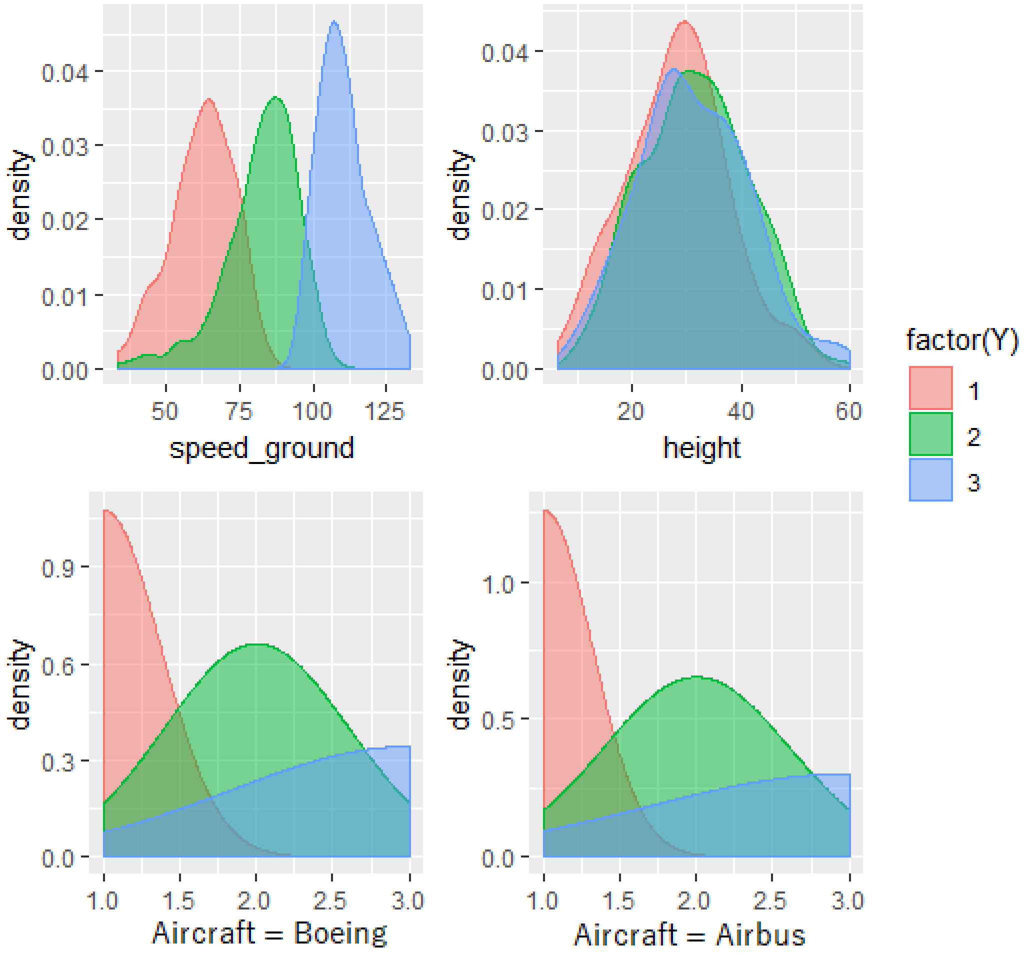


Table & Figure: 3 Visualization association of significant variables

Conclusion

The variables which are individually significant are Aircraft, Speed Ground, Speed Air & Height. However, since we are aware that Speed Air is highly correlated with Speed Ground, we will drop it from our model as well. Another advantage of doing so is that it has few NAs which cause errors during Multinomial Regression. We observe that Height and Aircraft although individually significant, do not contain much discriminatory power.

Step 3. Create model based on most significant variables

Code

```
#Model using significant variables
manual.model <- multinom(Y ~ speed_ground + height + aircraft,data = faa)
summary(manual.model)
z.manual <- summary(manual.model)$coefficients/summary(manual.model)$standard.errors
# 2-tailed Wald z tests to test significance of coefficients
p.manual <- (1 - pnorm(abs(z.manual), 0, 1)) * 2
```

Output

```
> summary(manual.model)
Call: multinom(formula = Y ~ speed_ground + height + aircraft, data = faa)

Coefficients:
  (Intercept) speed_ground    height aircraftboeing
2   -23.28484    0.2472743 0.1467859       3.982905
3  -126.43265    1.1756019 0.3782799       9.040905

Std. Errors:
  (Intercept) speed_ground     height aircraftboeing
2  1.88720542   0.01980816 0.01714538      0.4027433
3  0.04519312   0.01276020 0.03604886      0.7502719

Residual Deviance: 430.9527
AIC: 446.9527

> p.manual
  (Intercept) speed_ground height aircraftboeing
2           0            0      0              0
3           0            0      0              0
```

Observation

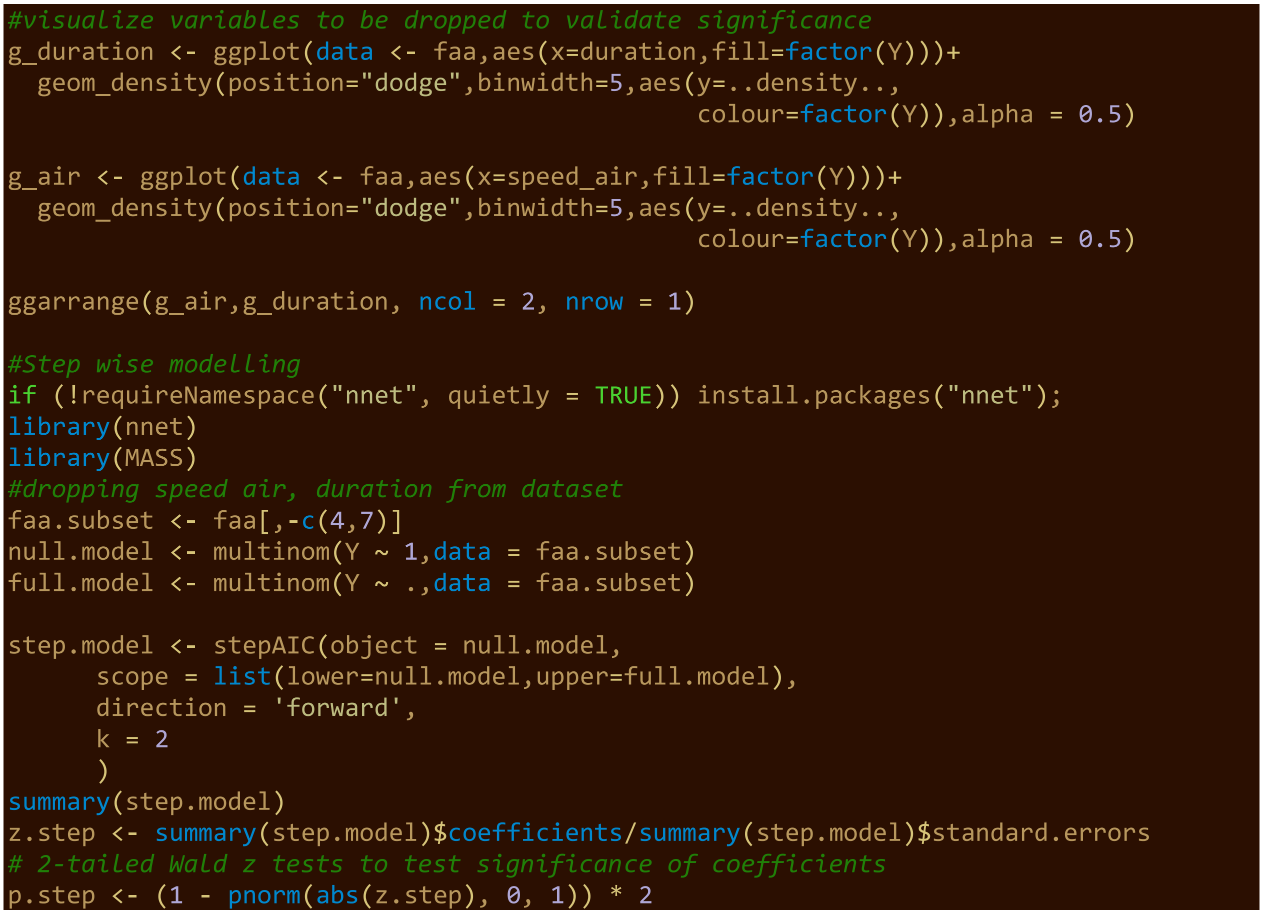
All p values are significant, and Height has less Coefficient size as compared to other Predictors.

Conclusion

We interpret the model in terms of log odds. For example, a unit increase in Speed Ground increases the log odds of Y being ‘2’ by 0.24 i.e. its coefficient size. The coefficient sizes observed in this step are coherent with our deductions from the previous step. The factors on which this model is built are ***Speed Ground, height & Aircraft***

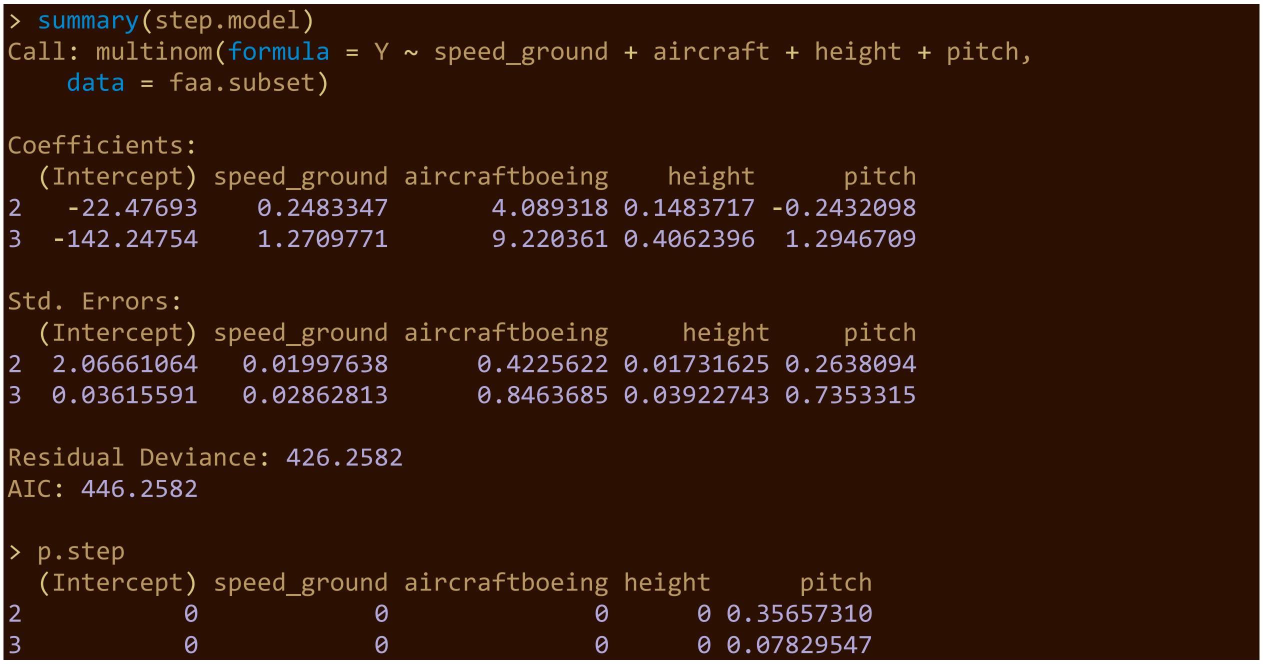
Step 4. Model selection using automated Forward Step AIC Function

Code



```
#visualize variables to be dropped to validate significance
g_duration <- ggplot(data <- faa,aes(x=duration,fill=factor(Y)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                                               colour=factor(Y)),alpha = 0.5)

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

ggarrange(g_air,g_duration, ncol = 2, nrow = 1)

#Step wise modelling
if (!requireNamespace("nnet", quietly = TRUE)) install.packages("nnet");
library(nnet)
library(MASS)
#dropping speed air, duration from dataset
faa.subset <- faa[,-c(4,7)]
null.model <- multinom(Y ~ 1,data = faa.subset)
full.model <- multinom(Y ~ .,data = faa.subset)

step.model <- stepAIC(object = null.model,
      scope = list(lower=null.model,upper=full.model),
      direction = 'forward',
      k = 2
      )
summary(step.model)
z.step <- summary(step.model)$coefficients/summary(step.model)$standard.errors
# 2-tailed Wald z tests to test significance of coefficients
p.step <- (1 - pnorm(abs(z.step), 0, 1)) * 2
```

Output



```
> summary(step.model)
Call: multinom(formula = Y ~ speed_ground + aircraft + height + pitch,
    data = faa.subset)

Coefficients:
  (Intercept) speed_ground aircraftboeing    height      pitch
2   -22.47693    0.2483347       4.089318 0.1483717 -0.2432098
3  -142.24754    1.2709771       9.220361 0.4062396  1.2946709

Std. Errors:
  (Intercept) speed_ground aircraftboeing     height     pitch
2  2.06661064   0.01997638      0.4225622 0.01731625 0.2638094
3  0.03615591   0.02862813      0.8463685 0.03922743 0.7353315

Residual Deviance: 426.2582
AIC: 446.2582

> p.step
  (Intercept) speed_ground aircraftboeing height      pitch
2           0            0              0      0 0.35657310
3           0            0              0      0 0.07829547
```

Observation

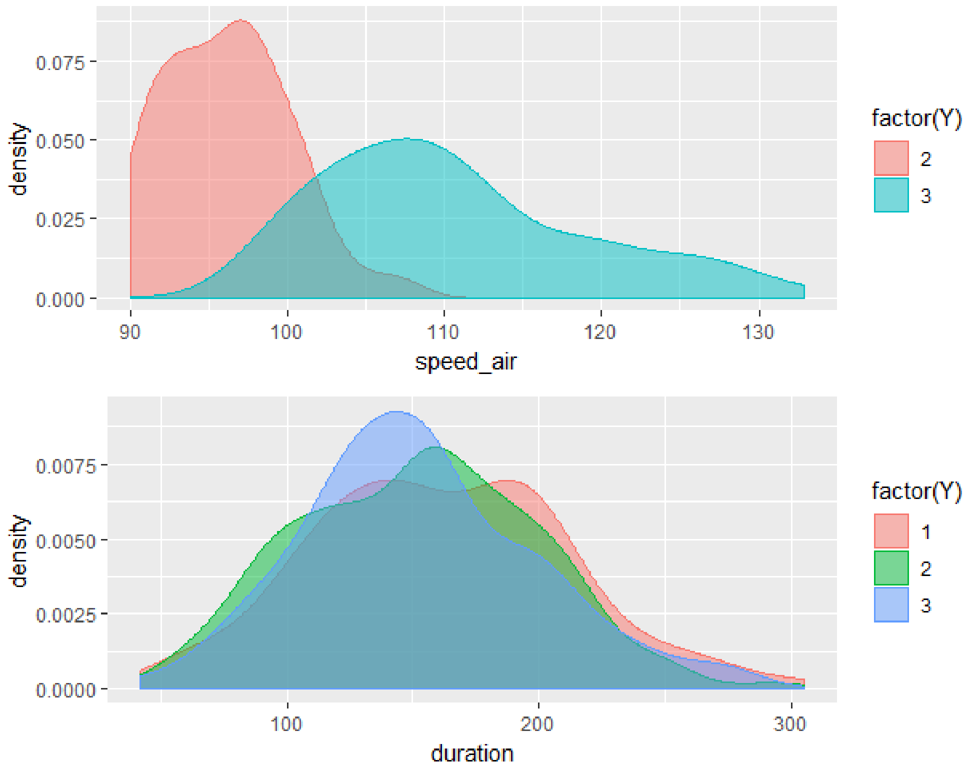


Table & Figure: 4 Distribution of Y by Speed Air & Duration

Note – We observe that Duration is not a satisfactory discriminatory factor for different levels of Y. Also, it has few missing values. In order to prevent any problems caused by this during Step Forward variable selection, we have removed it from the dataset. Similarly, we do know already that Speed Air is correlated with Speed Ground and it has few missing values. In the same fashion, it has been removed from the dataset.

Conclusion

We have performed Step Forward Variable Selection technique using AIC as criteria. The Null Model had no variables, only intercepts. The Full Model had all factors excluding Duration & Speed Air. The model selected by the algorithm contains ***Speed Ground, height, Pitch & Aircraft.*** This is slightly different from what we had for our manual model because Pitch is now selected. However, looking at the p values for coefficients of Pitch, we still see that it is statistically not significant

Step 5. Comparing models

Code

```
#comparing models check if aic is less
AIC(step.model)
AIC(manual.model)
AIC(step.model) < AIC(manual.model)

#checking deviance
deviance(manual.model)
deviance(step.model)

#calculate Chi square statistic
delta.dev <- deviance(step.model)-deviance(manual.model)
delta.degf <- step.model$edf-manual.model$edf
pchisq(delta.dev,abs(delta.degf),lower=F)
```

Output

```
> #comparing models check if aic is less
> AIC(step.model)
[1] 446.2582
> AIC(manual.model)
[1] 446.9527
> AIC(step.model) < AIC(manual.model)
[1] TRUE
> #checking deviance
> deviance(manual.model)
[1] 430.9527
> deviance(step.model)
[1] 426.2582
> #calculate Chi square statistic
> delta.dev <- deviance(step.model)-deviance(manual.model)
> delta.degf <- step.model$edf-manual.model$edf
> pchisq(delta.dev,abs(delta.degf),lower=F)
[1] 1
```

Observation

Model suggested by Step function is slightly better than the manually selected model in terms of AIC and Model Deviance. The difference is statistically insignificant as suggested by the Chi square test.

Conclusion

We have finalized the model suggested by Step Forward AIC with final factors as ***Speed Ground, height, Pitch & Aircraft.***

Step 6. In-sample Predictive power of the model

Code

```
#in sample predictions
y.pred <- predict(step.model,faa.subset)
y.prob <- predict(step.model,faa.subset,type='probs')

faa.subset$y.pred <- y.pred
faa.subset$y.prob <- y.prob


#visualize correlation with important factors
library(ggpubr)
library(ggplot2)

p_ground <- ggplot(data <- faa.subset,aes(x=speed_ground,fill=factor(y.pred)))+
  geom_density(aes(y=..density..,colour=factor(y.pred)),alpha = 0.5)

p_height <- ggplot(data <- faa.subset,aes(x=height,fill=factor(y.pred)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                                       colour=factor(y.pred)),alpha = 0.5)

p_pitch <- ggplot(data <- faa.subset,aes(x=pitch,fill=factor(y.pred)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                                       colour=factor(y.pred)),alpha = 0.5)

beoing.index <- which(faa$aircraft=='boeing')

p_aircraft.b <- ggplot(data <- faa.subset[-beoing.index,],
                       aes(x=factor(y.pred),fill=factor(y.pred)))+
                geom_density(position="dodge",
                aes(y=..density..,colour=factor(y.pred),alpha = 0.1))

p_aircraft.a <- ggplot(data <- faa.subset[-beoing.index,],
                aes(x=y.pred[-beoing.index],fill=factor(y.pred[-beoing.index])))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
               colour=factor(y.pred[-beoing.index])),alpha = 0.5)

ggarrange(p_ground,p_height,p_pitch,p_aircraft.b,
          p_aircraft.a, ncol = 2, nrow = 3)

#Confusion Matrix
xtabs(~faa.subset$y.pred+faa.subset$Y)
```

Output

```
> xtabs(~faa.subset$y.pred+faa.subset$Y)
                 faa.subset$Y
faa.subset$y.pred   1   2   3
                1 232  35   0
                2  37 419   6
                3   0   5  97
```

Observation

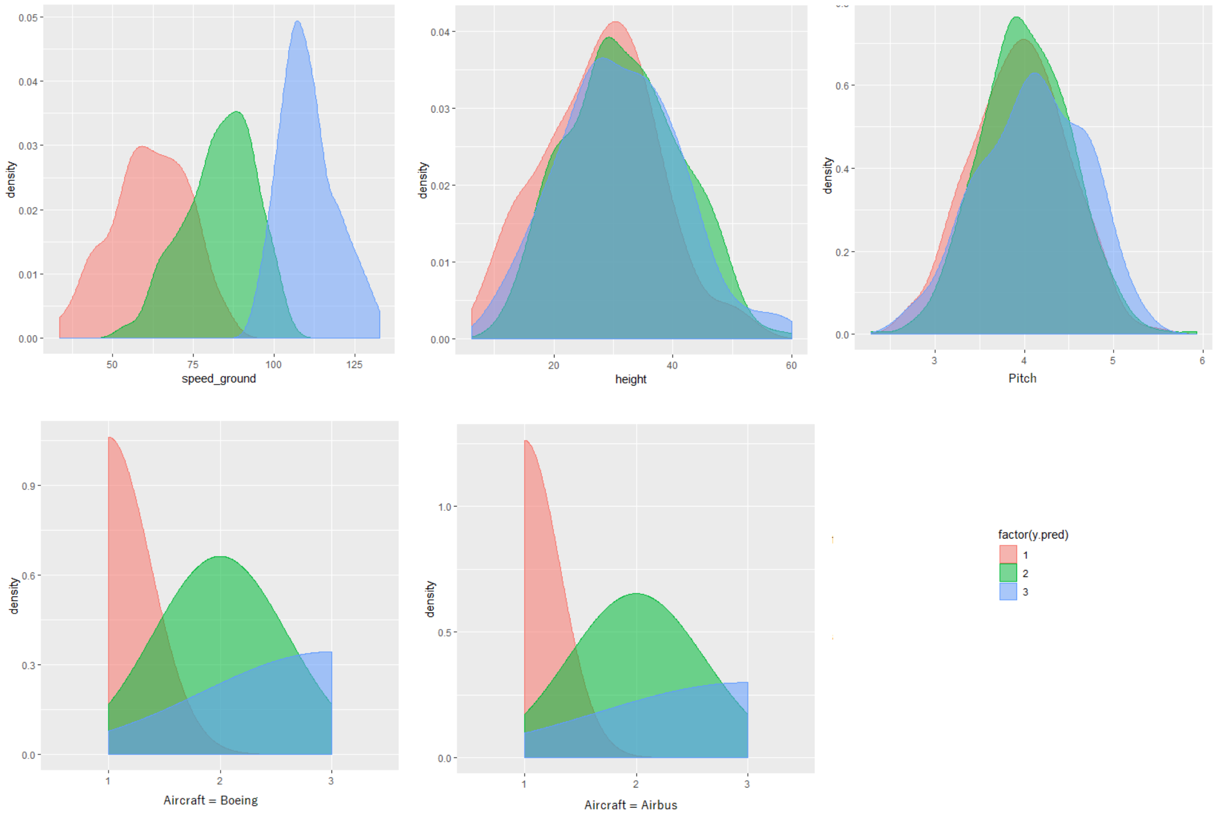


Table & Figure: 5 Prediction of Y by Significant Factors

Conclusion

The Misclassification Rate is 9.98. It means that the Model is misclassifying 10% of the time but is 90% accurate.

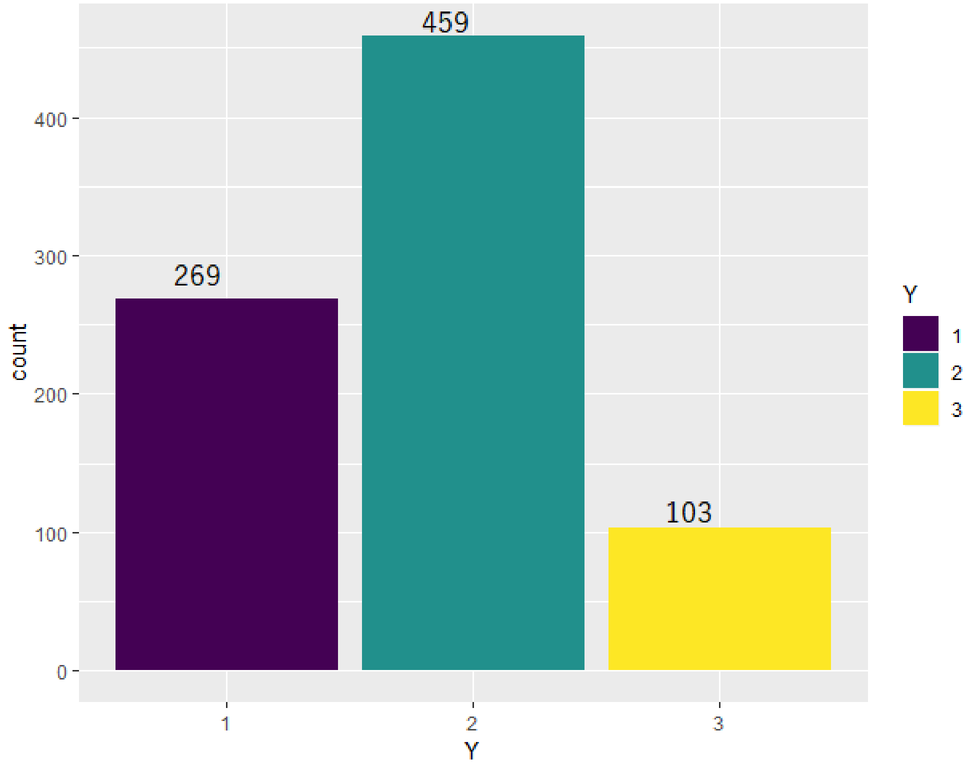
Step 7. Client presentation

Definition – Y is defined as a classifier for Landing distance based on the following conditions

Y = 1 if distance < 1000

Y = 2 if 1000 < = distance < 2500

Y = 3 otherwise



Model – Y is significantly impacted by *Speed Ground, height, Pitch & Aircraft.* The model can be interpreted in terms of Odds as follows. For instance, we interpret this as, '*Unit increase in the Speed Ground will increase the Odds of Landing Distance to be between 1000 m & 2500 m by 28%'*. *Pitch* may not be directly interpreted like other factors because individually, it is not significant.

Variables	2			3		
	Coefficients	Std Err	Odds %	Coefficients	Std Err	Odds %
(Intercept)	-22.47693	2.06661064	100	-142.24754	0.03615591	-100
speed_ground	0.2483347	0.01997638	28	1.2709771	0.02862813	256
aircraftboeing	4.089318	0.4225622	-5870	9.220361	0.8463685	1009971
height	0.1483717	0.01731625	16	0.4062396	0.03922743	50
pitch	-0.2432098	0.2638094	-22	1.2946709	0.7353315	265

Table & Figure: 6 Results of Final Model

Prediction Accuracy – The model can predict the correct class of landing distance with 90% accuracy. The highest rate of misclassification occurs on the lower side of the spectrum which further reduces the impact of a wrong prediction.

Graph – Following graph visualizes the relationship between different important factors and classes of Landing distance

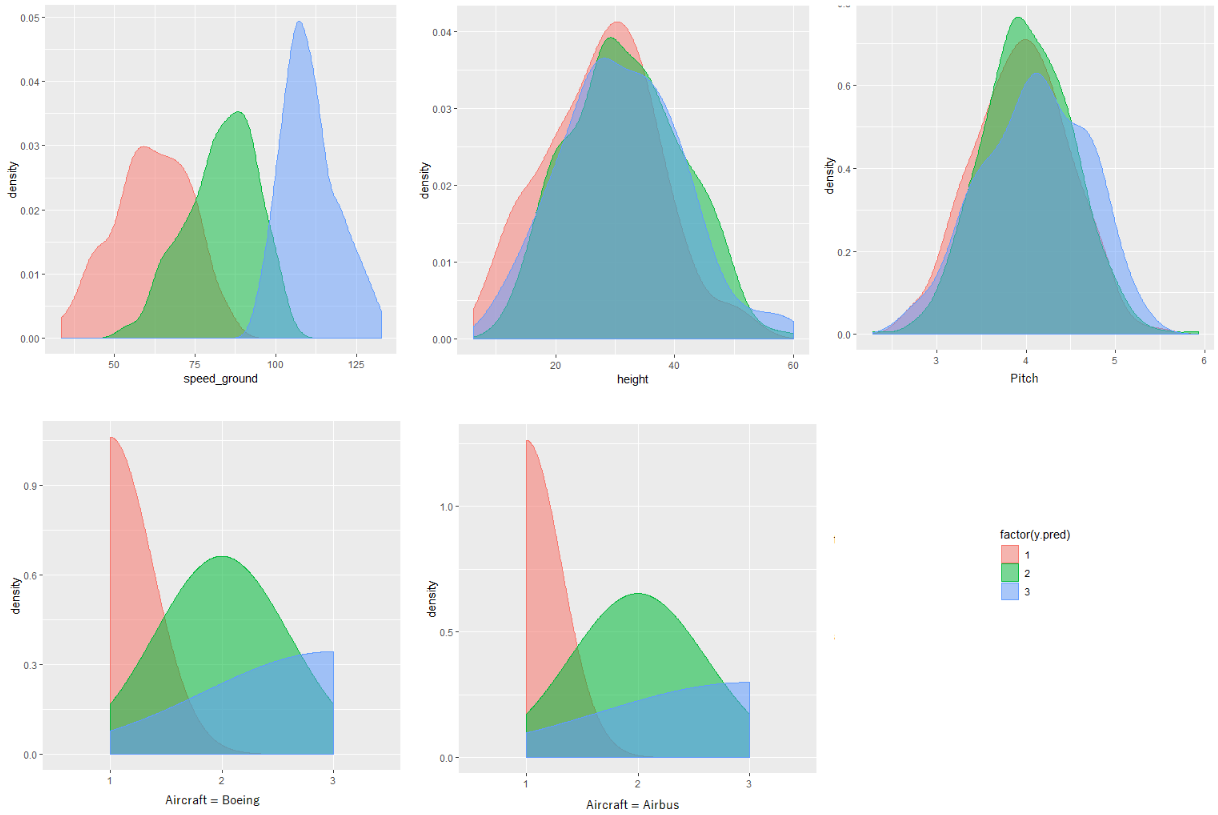


Table & Figure: 7 Distribution of Y by significant factors

Key Takeaways –

- Speed Ground, height, Pitch & Aircraft are significant factors
- Most significant factor based on various considerations is Speed Ground
- Misclassification Rate is 9.98%
- Mostly, wrong predictions by the model were on the lower side of the spectrum

Question 2. Modeling Count Data

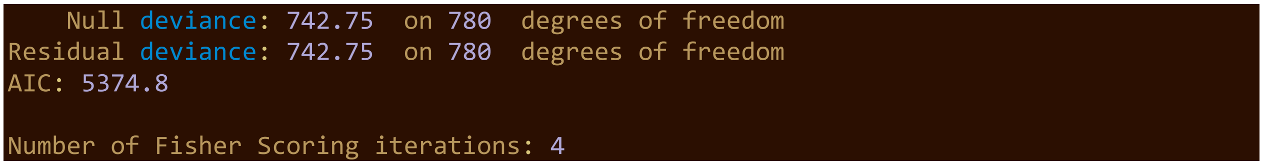
The number of passengers is often of interest of airlines. What distribution would you use to model this variable? Do we have any variables that are useful for predicting the number of passengers on board?

Code

```
plot(density(faa$no_pasg),main = 'Number of Passengers')
hist(faa$no_pasg,main = 'Number of Passengers')


scatterplotMatrix(~ no_pasg + Y + no_pasg + speed_ground +
                    speed_air + height + pitch + duration, data <- faa,
                  regLine = F, ellipse = F, diagonal = F,smooth = F  )


g_no_pasgn_dist <- ggplot(data <- faa,aes(x=no_pasg,fill=factor(aircraft)))+
  geom_density(position="dodge",aes(y=..density..,
                                    colour=factor(aircraft)),alpha = 0.5)
faa.np <- faa[,-4] %>% na.omit
null.model <- glm(no_pasg ~ 1,family='poisson',data = faa.np)
full.model <- glm(no_pasg ~ .,family='poisson',data = faa.np)

step.model <- step(object = null.model,
                      scope = list(lower=null.model,upper=full.model),
                      direction = 'forward',
                      k = 2)
summary(full.model)
summary(step.model)
```

Output

```
> summary(full_model)
Call: glm(formula = no_pasg ~ ., family = "poisson", data = faa.np)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.099e+00  4.934e-02  83.068   <2e-16 ***
aircraftboeing -2.259e-03  1.089e-02  -0.207    0.836
speed_ground    3.232e-04  4.432e-04   0.729    0.466
height          6.360e-04  5.032e-04   1.264    0.206
pitch          -1.968e-03  9.511e-03  -0.207    0.836
duration       -1.034e-04  9.594e-05  -1.078    0.281
Y              -1.264e-02  1.356e-02  -0.932    0.351

> summary(step.model)

Call: glm(formula = no_pasg ~ 1, family = "poisson", data = faa.np)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.095709   0.004616   887.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
```



```
    Null deviance: 742.75  on 780  degrees of freedom
Residual deviance: 742.75  on 780  degrees of freedom
AIC: 5374.8

Number of Fisher Scoring iterations: 4
```

Observation

First, we observe the distribution of the variable of interest. It looks like Normal density plot. However, we are aware that it cannot take decimal values and hence we will be using Poisson approximation instead.

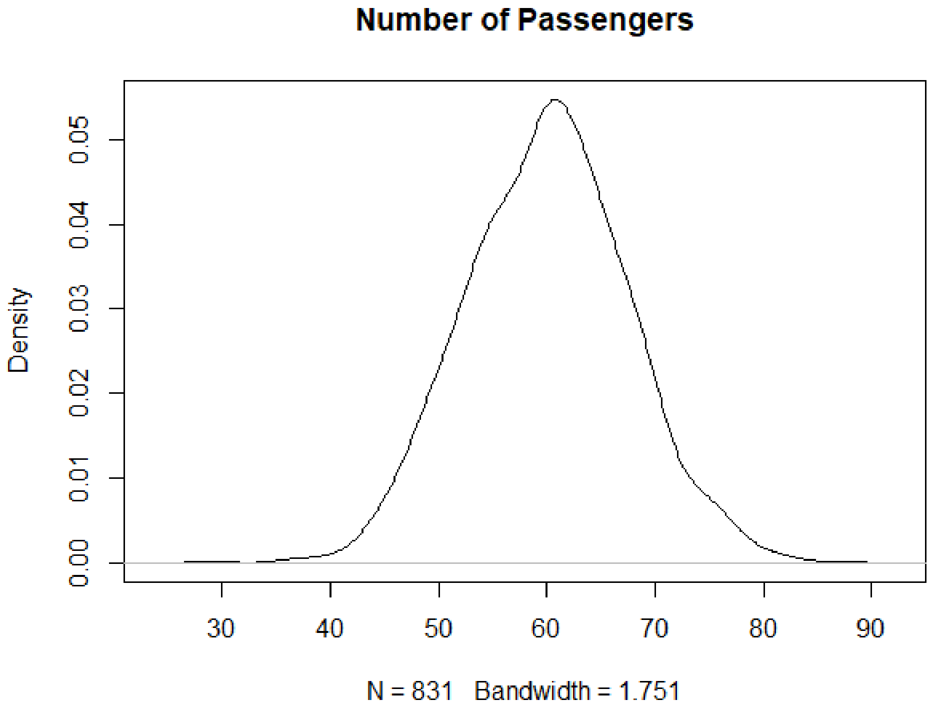


Table & Figure: 8 Distribution of Number of Passengers

In order to study the distribution in a better way, we plot the histogram –

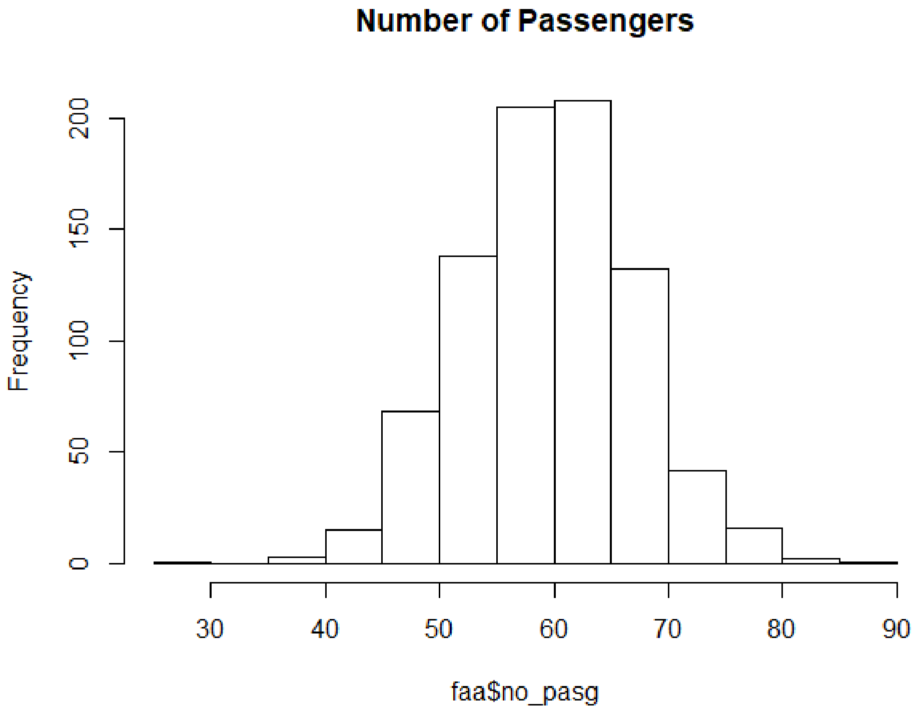


Table & Figure: 9 Histogram of Number of Passenger

We try to visualize if there are any significant variables which can show trends. However, as we observe from the below graph, that does not seem to be the case.

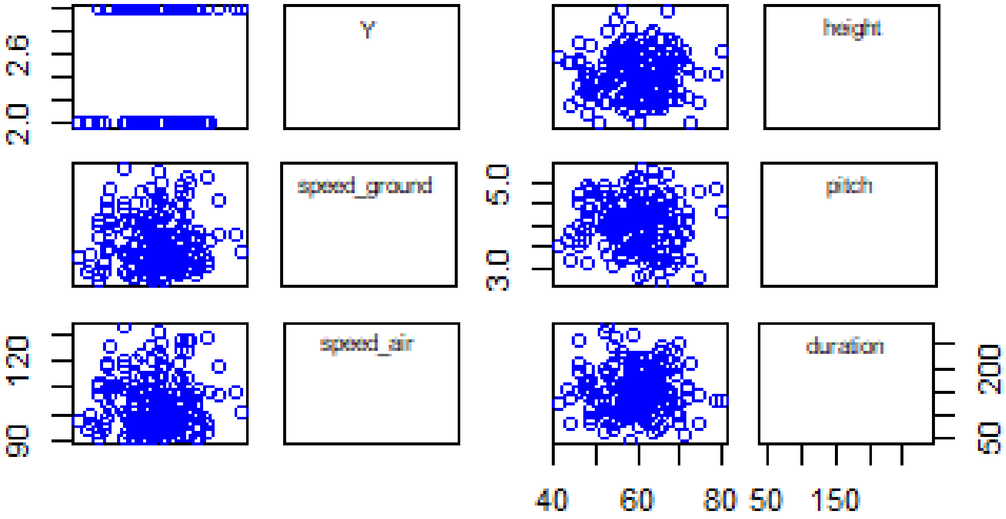


Table & Figure: 10 Number of Passengers versus all others

Technically, we are aware that Boeing and Airbus makes have different passenger capacities. We try to observe if there is any difference between the distribution of number of passengers between the two aircraft makes

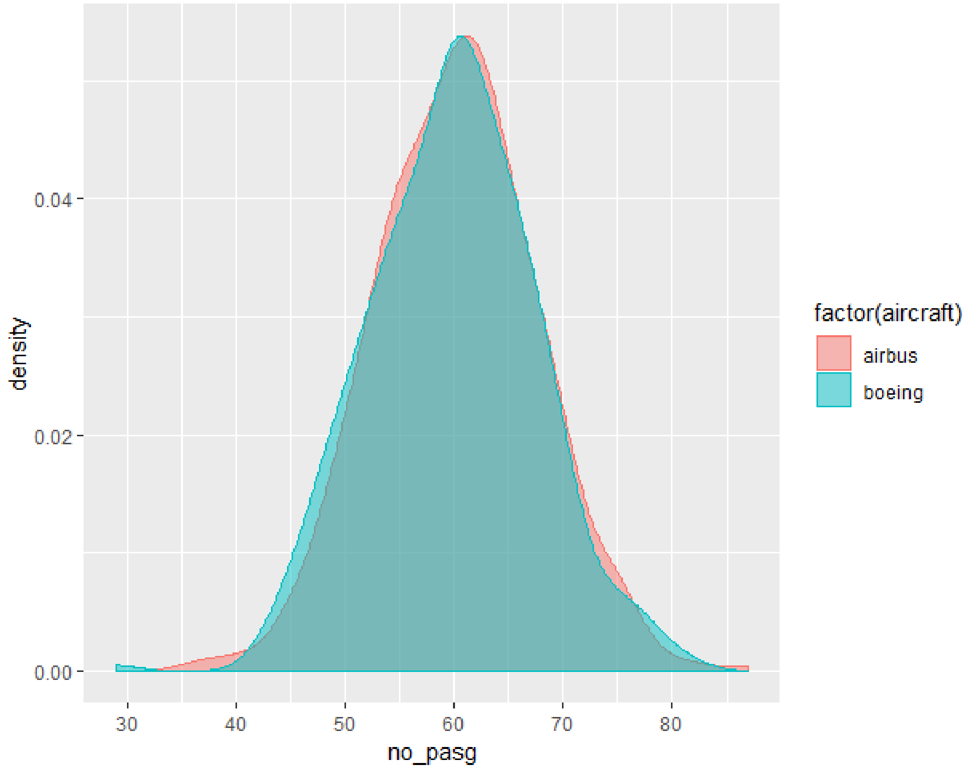


Table & Figure: 11 No. of Passenger by Aircraft

Conclusion

There seems to be no significant variable which can explain No. of Passengers. We ran a full model with all variables except Speed Air and none of them had a p_value less than 0.05. We also performed Step Forward selection using AIC as criteria, however, the final model suggested just had the intercept. Hence we conclude that no variable is correlated enough to explain the variability with number of Passengers and below is how the Fitted vs. Actual graph looks like a straight line, which should instead have looked like a diagonal line across the graph

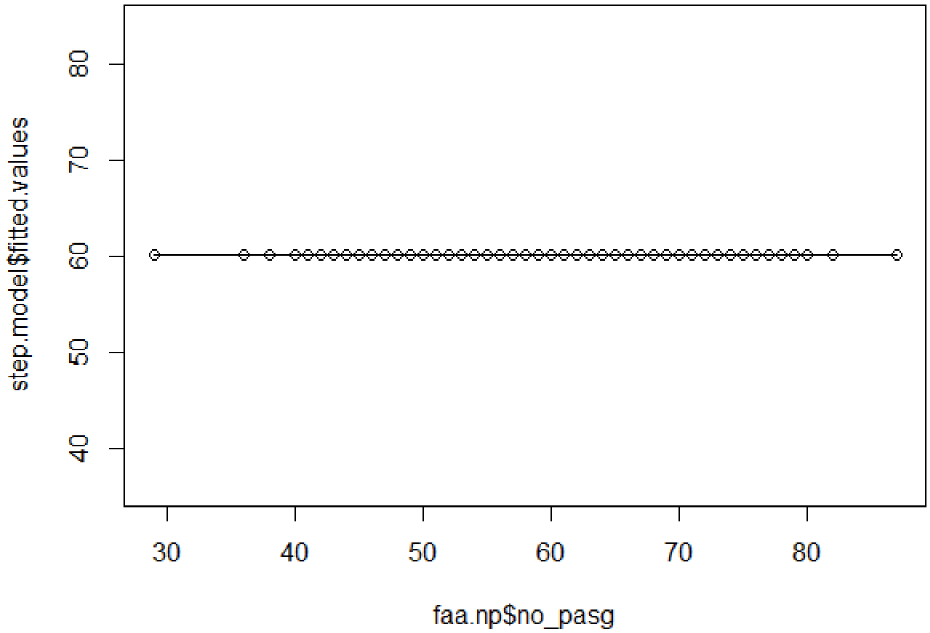


Table & Figure: 12 Fitted vs Actual for No. of Passenger