Statistical Modeling Project 3

Ashwita Saxena
M06119969

Q1. 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!

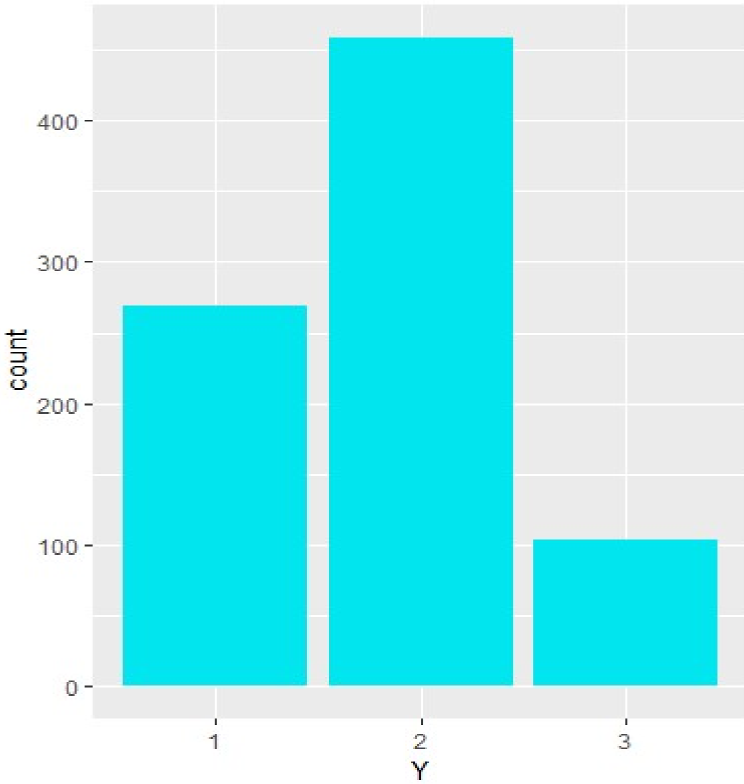
CREATING MULTINOMIAL VARIABLE FOR DISTANCE

```
# importing data -----------------------------------------------------------

library(readxl)
FAA1<-read_xlsx("D:/MSBA/Spring Sem/statistical modeling/FAA1(1).xlsx",)
FAA2<-read_xlsx("D:/MSBA/Spring Sem/statistical modeling/FAA2(1).xlsx",)

#### structure
str(FAA1)
str(FAA2)

#### merging
library(dplyr)
FAA1_2 <- select(FAA1, aircraft, no_pasg, speed_ground, speed_air, height, pitch,
distance)

merged<- rbind(FAA1_2,FAA2)

sum(duplicated(merged))

new <- unique(merged)
summary(new)

#adding duration back in

final <- left_join(new, FAA1)

#### checking the final dataset
str(final)
summary(final)
#standard deviation
sd_vector <- c("distance","duration","no_pasg", "height", "speed_ground", "speed_air",
"pitch")
sapply(final[sd_vector], sd, na.rm=T)

# data cleaning -----------------------------------------------------------

a <- filter(final, speed_ground >= 30, speed_ground <= 140, height >= 6, distance <= 6000
)
#remove duration <40 but keep NAs
```

Statistical Modeling Part 3
Ashwita Saxena
M06119969

```
clean <- a[(a$duration >= 40 | is.na(a$duration)),]

summary(clean)

data <- clean

data$aircraft <- as.factor(data$aircraft)

str(data)
summary(data)
#standard deviation
sd_vector <- c("distance","duration","no_pasg", "height", "speed_ground", "speed_air",
"pitch")
sapply(data[sd_vector], sd, na.rm=T)

hist(data$distance)
boxplot(data$distance)

# Discretization of landing distance (creating multinomial variable )--------------------
--------------

for (i in 1:831) {
  if (data$distance[i] < 1000) {
    data$distance[i] = 1
  } else if (data$distance[i] >= 1000 & data$distance[i] < 2500){
    data$distance[i] = 2
  } else data$distance[i] = 3
}

table(data$distance)

data$distance <- as.factor(data$distance)

names(data)[7] <- "Y"

str(data)

## visualization

par(mfrow=c(1,1))

library(ggplot2)
ggplot(data = data, aes(x = Y)) +
  geom_bar(stat ="count", fill = 'turquoise2')
```

```
> table(data$distance)

  1   2   3 
269 459 103 
```



Observation: Our final dataset has 831 observations and 8 variables. Y is the distance variable that has been classified into a multinomial variable with values 1,2,3. There are 269 observations where Y is 1, 459 observations where Y is 2, and 103 observations where Y is 3. This can be shown in the bar plot above.

Conclusion: Total number of observation in the dataset are 831 and it contains 8 variables.

SIGNIFICANCE OF INDIVIDUAL MODELS

```
# significance of individual models ---------------------------------------

### single variable models
library(nnet)

## speed_ground
model1 <- multinom (Y ~ speed_ground,data)
summary(model1)
sig_CI_1 <- c(summary(model1)$coefficients[,2]-
1.96*(summary(model1)$standard.errors[,2]),
summary(model1)$coefficients[,2]+1.96*(summary(model1)$standard.errors[,2]))
sig_CI_1

## height
model2 <- multinom (Y ~ height,data)
summary(model2)
sig_CI_2 <- c(summary(model2)$coefficients[,2]-
1.96*(summary(model2)$standard.errors[,2]),
summary(model2)$coefficients[,2]+1.96*(summary(model2)$standard.errors[,2]))
sig_CI_2
```

```
## duration
model3 <- multinom (Y ~ duration,data)
summary(model3)
sig_CI_3 <- c(summary(model3)$coefficients[,2]-
1.96*(summary(model3)$standard.errors[,2]),
summary(model3)$coefficients[,2]+1.96*(summary(model3)$standard.errors[,2]))
sig_CI_3

## pitch
model4 <- multinom (Y ~ pitch,data)
summary(model4)
sig_CI_4 <- c(summary(model4)$coefficients[,2]-
1.96*(summary(model4)$standard.errors[,2]),
summary(model4)$coefficients[,2]+1.96*(summary(model4)$standard.errors[,2]))
sig_CI_4

## aircraft
model5 <- multinom (Y ~ aircraft,data)
summary(model5)
sig_CI_5 <- c(summary(model5)$coefficients[,2]-
1.96*(summary(model5)$standard.errors[,2]),
summary(model5)$coefficients[,2]+1.96*(summary(model5)$standard.errors[,2]))
sig_CI_5

## speed air
model6 <- multinom (Y ~ speed_air,data)
summary(model6)
sig_CI_6 <- c(summary(model6)$coefficients[2]-1.96*(summary(model6)$standard.errors[2]),
summary(model6)$coefficients[2]+1.96*(summary(model6)$standard.errors[2]))
sig_CI_6


## no_pasg
model7 <- multinom (Y ~ no_pasg,data)
summary(model7)
sig_CI_7 <- c(summary(model7)$coefficients[,2]-
1.96*(summary(model7)$standard.errors[,2]),
summary(model7)$coefficients[,2]+1.96*(summary(model7)$standard.errors[,2]))
sig_CI_7
```

```
> sig_CI_1
        2         3         2         3 
0.1495743 0.7246769 0.1118557 0.4763725 
> sig_CI_2
          2           3           2           3 
0.021829852 0.009210762 0.053978834 0.056665013 
> sig_CI_3
            2             3             2             3 
-0.0071353220 -0.0083873671 -0.0006258771  0.0012943951 
> sig_CI_4
          2           3           2           3 
-0.11050481  0.07732464  0.46450876  0.94833262 
> sig_CI_5
        2         3         2         3 
0.5295954 0.9303285 1.1620364 1.8905767 
> sig_CI_6
speed_air speed_air 
0.3597662 0.6641312 
> sig_CI_7
          2           3           2           3 
-0.02438272 -0.04030084  0.01584606  0.02040191 
```

Observation: We observe that when we add and subtract 1.96 times the standard error to the coefficients of the model, we get a confidence interval. If that interval covers zero, the variable is insignificant. If that variable does not cover zero, that variable is significant. Based on this calculation, we can say the significance of the variables looks like the following

Variable	Significance
speed_ground	significant
height	significant
duration	not significant for one level of multinomial variable
pitch	not significant for one level of multinomial variable
aircraftboeing	significant
speed_air	significant
no_pasg	not significant for one level of multinomial variable

Summary of model 1:

```
multinom(formula = Y ~ speed_ground, data = data)

Coefficients:
  (Intercept) speed_ground
2   -9.001649    0.1307150
3  -56.689405    0.6005247

Std. Errors:
  (Intercept) speed_ground
2   0.6985136  0.009622096
3   6.3343700  0.063342960

Residual Deviance: 723.9604
AIC: 731.9604
```

To interpret the coefficient of speed ground when Y = 2: as speed ground increases by one unit, the log odds of landing distance being in category Y=2 vs not being in Y=2 increase by 0.13. Similarly, other coefficients can be interpreted

Conclusion: We see that speed ground, speed air, aircraft type boeing and height are significant variables for our multinomial response Y. For further analysis, we will not use speed_air as it is highly correlated with speed_air.

MANUAL MODEL WITH SIGNIFICANT VARIABLES

```
# create multivariate model manually ------------------------------------------------

model_manual <- multinom(Y ~ speed_ground + height + aircraft, data=data)
summary(model_manual)
```

```
sig_CI_manual <- data.frame(summary(model_manual)$coefficients[,2:4]-
1.96*(summary(model_manual)$standard.errors[,2:4]),
summary(model_manual)$coefficients[,2:4]+1.96*(summary(model_manual)$standard.errors[,2:4
]))
sig_CI_manual
```

Output:

```
> summary(model_manual)
Call:
multinom(formula = Y ~ speed_ground + height + aircraft, data = data)

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


>sig_CI_manual
  speed_ground    height aircraftboeing speed_ground.1  height.1 aircraftboeing.1
2    0.2084503 0.1131810       3.193528      0.2860983 0.1803909         4.772282
3    1.1505919 0.3076242       7.570372      1.2006119 0.4489357        10.511438
```

Observation:

We can see that none of the confidence intervals contain zero. That means all the variables in this model (speed ground, height and aircraft type boeing) are significant. Looking at the summary of the model, we can say that, with a unit increase in speed ground the log odds of landing distance being in category Y=2 vs not being in Y=2 increase by 0.24. In the same way, with a unit increase in speed ground, the log odds of landing distance being in category Y=3 vs not being in Y=3 increase by 1.15. We can interpret the other variables of the model similarly.

Conclusion: Final manual model shows speed ground, height and aircraft type boeing as significant

STEP AIC MODEL

```
# create automated variable selected model --------------------------------

## model selection based on AIC
data_step <- data[,-c(4,8)]
null_model <- multinom(Y ~ 1,data = data_step)
full_model <- multinom(Y ~ .,data = data_step)
```

```
model_step <- step(object = null_model,scope =
list(lower=null_model,upper=full_model),direction = 'forward',k = 2)
summary(model_step)

## model comparison based on significance

deviance(model_step) - deviance(model_manual)
model_step$edf - model_manual$edf
pchisq(deviance(model_step)-deviance(model_manual),-
model_manual$edf+model_step$edf,lower=F)
```

```
multinom(formula = Y ~ speed_ground + aircraft + height + pitch,
    data = data_step)

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
```

```
> deviance(model_step) - deviance(model_manual)
[1] -4.694494
> model_step$edf - model_manual$edf
[1] 2
> pchisq(deviance(model_step)-deviance(model_manual),-model_manual$edf+model_s
tep$edf,lower=F)
[1] 1
```

Observation: We see that in the step AIC model, pitch is included in the model, however it is not overall significant. The AIC of step AIC model is smaller than that of the manual model. The deviance of step AIC model is also smaller than that of the manual model. Hence we will select step AIC model as our final model.

Conclusion: final model is Y ~ speed_ground + aircraft + height + pitch

PREDICTION

```
# Prediction -----------------------------------------------------------

xtabs(~predict(model_step)+data$Y)

(35+37+5+6)/(232+419+97+35+37+5+6)
```

```
                   data$Y
predict(model_step)   1   2   3
                  1 232  35   0
                  2  37 419   6
                  3   0   5  97
> (35+37+5+6)/(232+419+97+35+37+5+6)
[1] 0.09987966
```

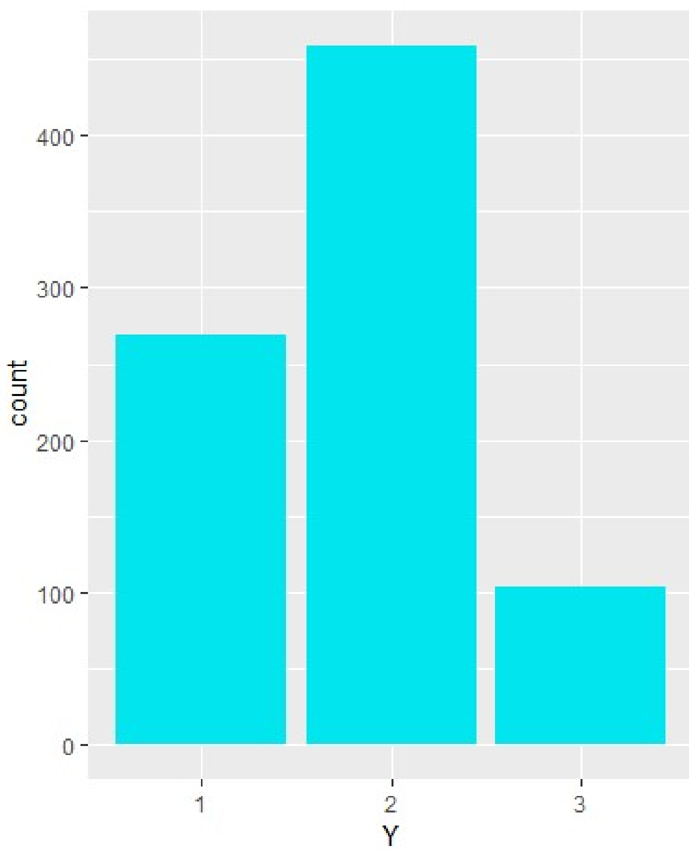
Observation: we obtain a confusion matrix from our final model and calculate the misclassification rate. We see that the misclassification rate of our model is 9.98% which is acceptable.

Conclusion: Misclassification Rate of our model is 9.98%.

PRESENTATION:

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

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

plot(jitter(as.numeric(Y),0.1)~jitter(as.numeric(aircraft)),data,xlab = "Airbus vs
Boeing",ylab="Y")
```

- We are calculating the impact of variables on different classes of landing distance. We classified the distance variable into three sections (Y variable). Y = 1 if distance < 1000, Y = 2 if 1000 < = distance < 2500, Y = 3 otherwise. There are 269 observations where distance is short (Y = 1), 459 observations where distance is medium (Y=2), and 103 observations where distance is long (Y = 3).



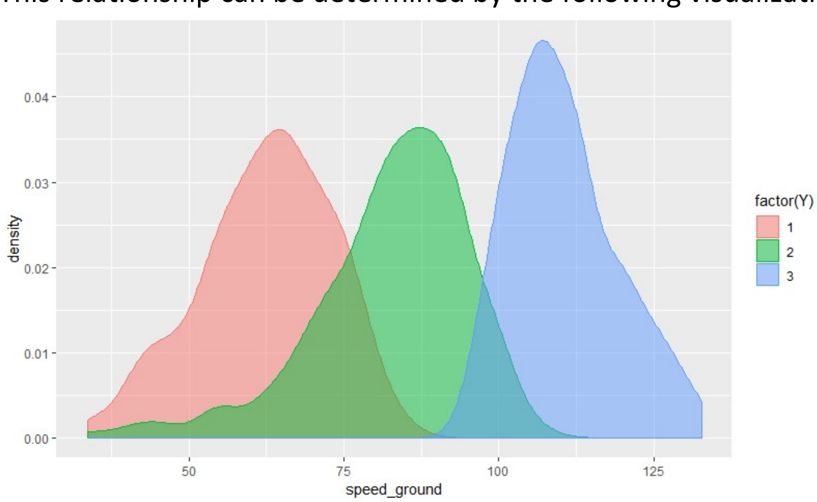
- Based on our analysis. We infer that ground speed, height of aircraft while passing over runway, and aircraft type boeing are the most important variables in predicting the multinomial variable Y.

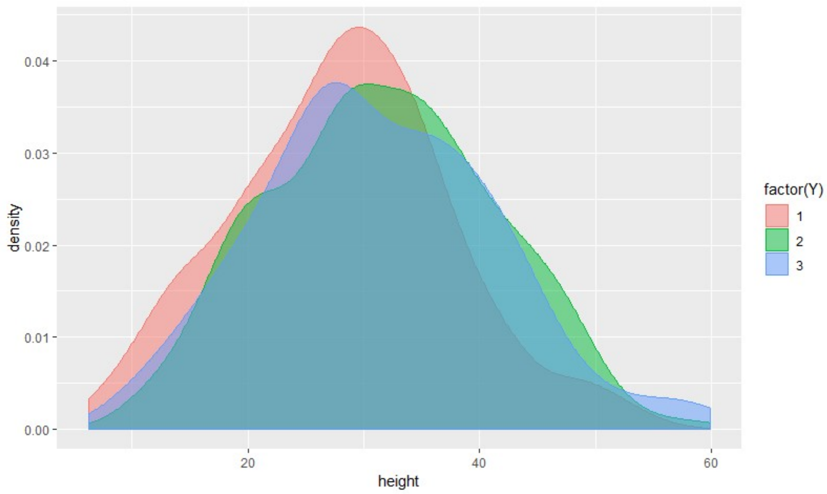
Final Model Y = B0 + B1(Speed Ground) + B2(aircraft) + B3(height) + B4(Pitch)

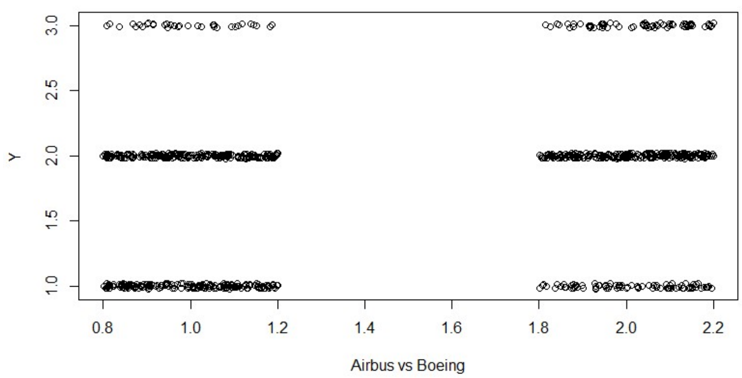
Variable	Class	Significance	Coefficient	Exponential of Beta
Speed ground	2	Yes	0.2483347	1.281888909
	3	Yes	1.2709771	3.564333572
aircraftboeing	2	Yes	4.089318	59.69916299
	3	Yes	9.220361	10100.71003
height	2	Yes	0.1483717	1.159943967
	3	Yes	0.4062396	1.501162188
pitch	2	No	-0.2432098	0.784106991
	3	No	1.2946709	3.649794629

- One mile per hour increase in ground speed increases the odds of medium distance (Y=2) as compared to short and long distance (Y ≠ 2) by 28%.
- One mile per hour increase in ground speed increases the odds of long distance (Y=3) as compared to short and medium distance (Y ≠ 3) by 256%.
- One meter increase in height increases the odds of medium distance (Y=2) as compared to short and long distance (Y ≠ 2) by 15.99%.
- One meter increase in height increases the odds of long distance (Y=3) as compared to short and medium distance (Y ≠ 3) by 50%.
- When aircraft type changes from airbus to boeing, the odds ratio for medium distance vs non medium distance is 59.69.
- When aircraft type changes from airbus to boeing, the odds ratio for long distance vs non long distance is 10100.
- We cannot interpret coefficients of pitch because we donot observe statistical significance

• This relationship can be determined by the following visualizations as well:



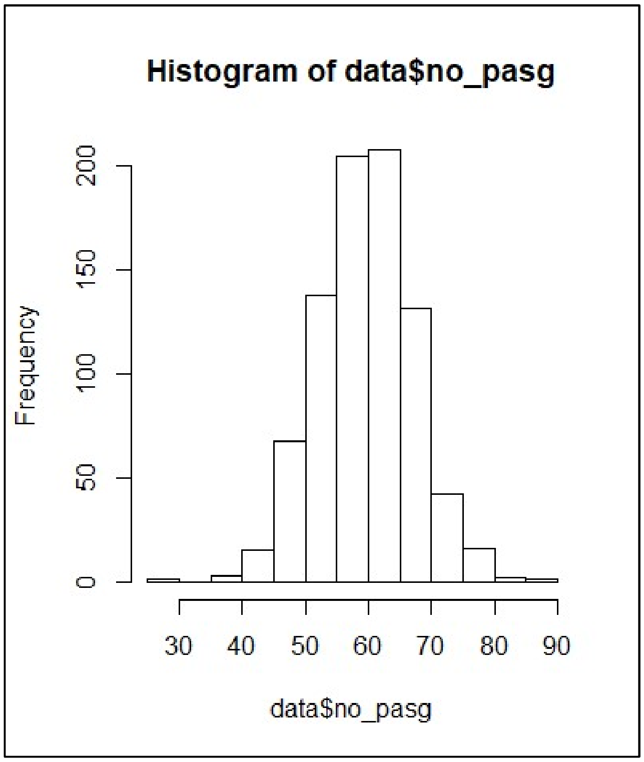




- Based on the 831 observations in the final dataset, our model predicts with 90% accuracy. If we have more data, we would be able to better predict impact of variables on different classes of distance.

Q2. 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?

Since number of passengers is count data, we will use Poisson distribution to model this variable. The distribution of this variable looks like the following




```
######################
# modeling count data #
######################

hist(data$no_pasg)

#removing speed air due to multicollinearity and null observations
data <- select(data, -speed_air) %>% na.omit()

model_poisson <- glm(no_pasg ~ . , family=poisson, data)
summary(model_poisson)
```

```
glm(formula = no_pasg ~ ., family = poisson, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4330  -0.6725   0.0328   0.6274   3.1643

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.076e+00  5.915e-02  68.922   <2e-16 ***
aircraftboeing -2.887e-04  1.185e-02  -0.024    0.981
speed_ground    5.150e-04  6.152e-04   0.837    0.402
height          6.612e-04  5.113e-04   1.293    0.196
pitch          -1.764e-03  9.516e-03  -0.185    0.853
distance       -1.259e-05  1.326e-05  -0.950    0.342
duration       -9.911e-05  9.587e-05  -1.034    0.301
---
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: 739.18  on 774  degrees of freedom
AIC: 5383.2

Number of Fisher Scoring iterations: 4
```

```
## variable selection using stepwise AIC
step_model_poisson <- step(model_poisson)
summary(step_model_poisson)
```

```
glm(formula = no_pasg ~ 1, family = poisson, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4627  -0.6652  -0.0106   0.6261   3.2525

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: Based on complete model as well as stepwise variable selection model, we can see that none of the variables are statistically significant in prediction of number of passengers.

Conclusion: Based on our data, we cannot predict number of passengers.