**Question 1**

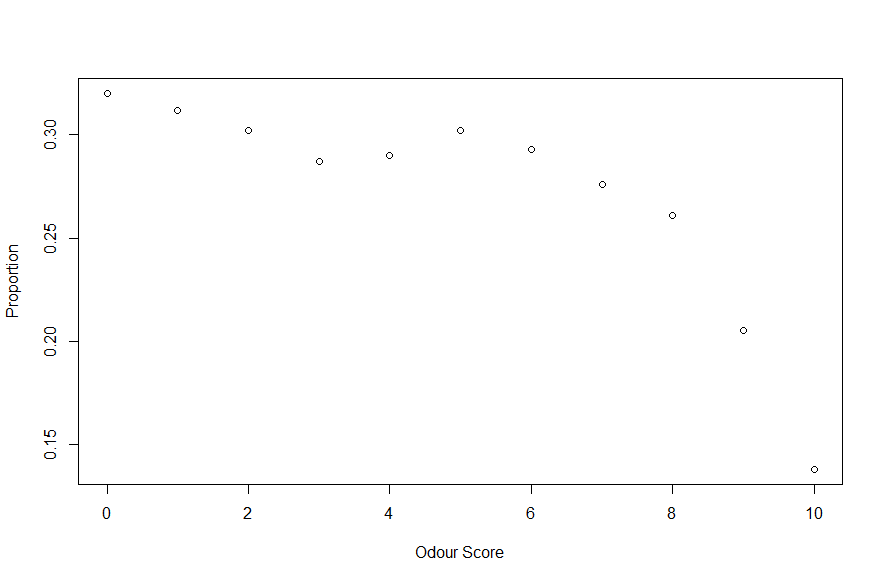


Figure 1: Odour score vs proportion

As the odour score increases, proportion of patients with dementia reduces

glm(formula = proportion ~ odour\_score, family = "binomial",

data = data\_odour, weights = total)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8754 -0.3170 -0.2202 0.7811 0.8686

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.62011 0.13601 -4.559 5.13e-06 \*\*\*

odour\_score -0.07599 0.02235 -3.401 0.000673 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 18.4855 on 10 degrees of freedom

Residual deviance: 6.8493 on 9 degrees of freedom

AIC: 62.991

Number of Fisher Scoring iterations: 4

The following inference can be drawn from the above output: The odour score has a negative effect on the proportion of people with dementia because it can be seen that its coefficient is negative.

* The deviance of the model without an intercept is given by the null deviance which is 18.4855
* The coefficient of odour score is significant (p = 0.000673) (coefficient of odour score is statistically different from 0)
* Odour Score is useful in predicting the proportion of patients with dementia
* With a negative coefficient β1, probability of getting patients with dementia decreases with increasing odour score
* The fitted equation is given by:

Estimate Std. Error 2.5 % 97.5 %

(Intercept) -0.620 0.136 -0.89 -0.356

odour\_score -0.076 0.022 -0.12 -0.032

From the 95% confidence interval output above:

* It can be seen that the 95% interval of the regression coefficient does not include "0" therefore, it is significant.
* Provided the model is used in regression, 95% of the time the coefficient of odour\_score will lie between -0.12 and -0.032.

Analysis of Deviance Table

Model: binomial, link: logit

Response: proportion

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 10 18.4855

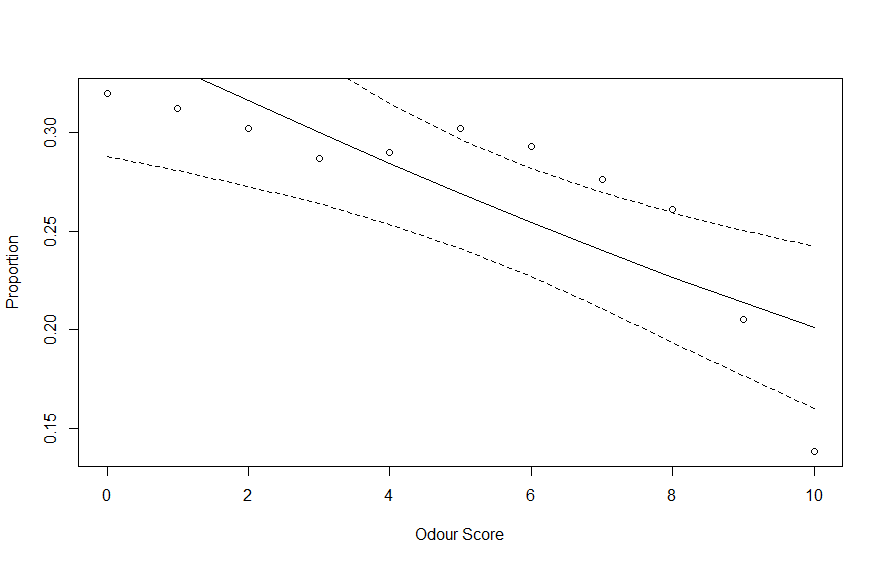
odour\_score 1 11.636 9 6.8493 0.0006468 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

From the above ANOVA output:

* It can be seen that the overall model is significant and the addition of the main term and interaction term is significant.
* The residual deviance of 6.8493, has an approximate chi-square distribution on 9 df and is not significant (p = 0.652806)
* The model provides adequate fit to the data because there is no significant difference between the model and observed data (p-value is above 0.05)
* The chi-square test associated with the residual deviance of the fitted model provided the p-value of 0.652806. Given that this value is greater than 0.05, the null hypothesis is not rejected. Therefore, the model using odour score as a predictor provides a good fit for the data.



The above figure shows the fitted model showing the 95% confidence interval for the binomial data showing the proportion of patients with dementia at various odour score. It can be seen that the model fits the observations adequately and more of the observations are captured within the 95% interval boundary.

Provided 25% of patients are diagnosed with dementia, the odour score would be calculated as:

**Question 2**

The Titanic dataset consist of a tally count of individuals that survived and total people present on the ship. The total count is grouped into different categories, one representing the class in which an individual can be found (crew, first, second, third), sex (male or female) and age (adult or children).

The first class is set as the reference level for the purpose of this analysis so as to properly compare the difference in survival rate of children in different class groups.

The figure below shows the percentage individuals that survived in different class groups. The chart illustrates that a larger number of crew members survives compared to other groups and the class with the least number of survivors are those in the third class. The figure also shows that adults consist of the majority of the total survivors, and lastly, there were almost as many females as male survivors as females consists of 48.4% of the population as compared to 51.6% of male.

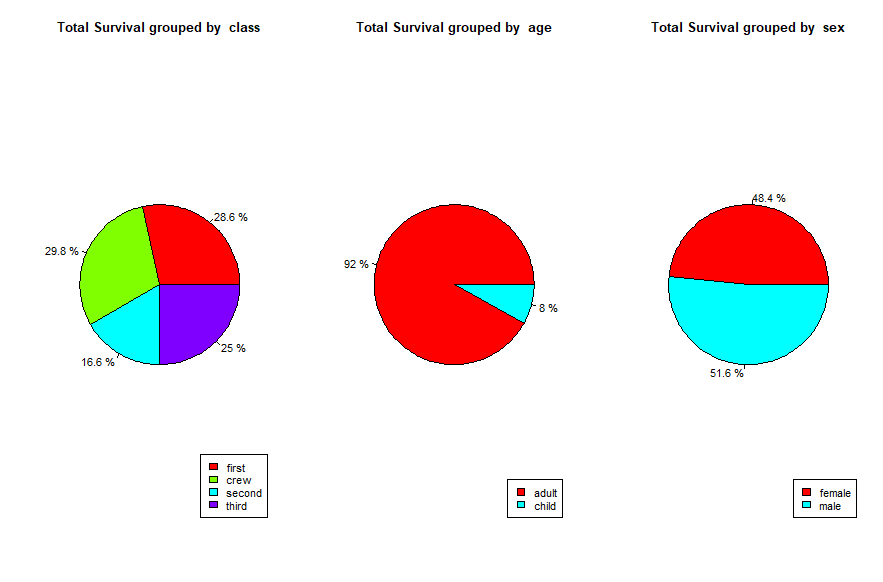


Figure 2: Total percentage survivors of each group

The bar chart shows the average survival rate of each group and it will be noticed that based on classification, individuals in the first-class have a higher survival rate on average and the third-class people has the least chance of surviving. Also, it is seen that on average children have the higher survival rate when compared to adults, lastly, provided an individual Is of the female sex, she has a higher survival rate when compared t male gender.

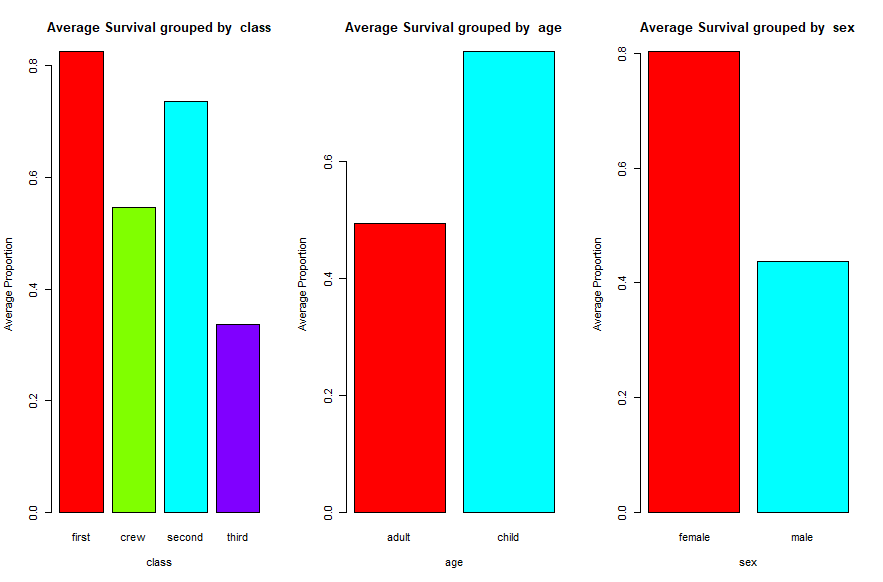


Figure 3: Average Survival rate

We determine the significant predictors and interactions using stepwise Logistic regression which tries to reduce the AIC value to the minimal by adding the relevant predictors to perform the regression

* Start: AIC=720.58

surv\_prop ~ 1

Df Deviance AIC

+ sex 1 237.45 288.07

+ class 3 491.22 545.84

+ age 1 652.38 703.00

<none> 671.96 720.58

* Step: AIC=288.11

surv\_prop ~ sex

Df Deviance AIC

+ class 3 131.20 187.86

+ age 1 231.54 284.21

<none> 237.45 288.11

* Step: AIC=188.04

surv\_prop ~ sex + class

Df Deviance AIC

+ class:sex 3 66.334 129.17

+ age 1 112.273 171.11

<none> 131.200 188.04

* Step: AIC=128.86

surv\_prop ~ sex + class + sex:class

Df Deviance AIC

+ age 1 45.928 110.45

<none> 66.334 128.86

* Step: AIC=110.52

surv\_prop ~ sex + class + age + sex:class

Df Deviance AIC

+ class:age 2 1.702 70.295

+ age:sex 1 37.258 103.852

<none> 45.928 110.522

* Step: AIC=70.31

surv\_prop ~ sex + class + age + sex:class + class:age

Df Deviance AIC

<none> 1.702 70.308

+ age:sex 1 0.000 70.606

A stepwise Logistic regression in the forward direction is used to determine the relevant predictors and interaction predictor for this model. And the result shows that all interaction predictors are significant except the interaction between age and sex which implies that the rate of survival cannot be explained by the age group (adult or child) of the different available sex (male or female).

glm(formula = surv\_prop ~ sex + class + age + sex:class + class:age,

family = "binomial", data = data\_titanic, weights = total)

Deviance Residuals:

1 2 3 4 5 6 7 8 9 10 11

0.00000 0.00000 0.00000 0.00005 0.00000 0.00000 0.00001 0.00007 0.00000 0.00000 -0.87878

12 13 14

0.83052 0.38247 -0.30603

Coefficients: (1 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.547e+00 5.051e-01 7.022 2.18e-12 \*\*\*

sexmale -4.273e+00 5.302e-01 -8.059 7.67e-16 \*\*\*

classcrew -1.646e+00 7.997e-01 -2.058 0.03955 \*

classsecond -1.732e+00 5.869e-01 -2.951 0.00317 \*\*

classthird -3.763e+00 5.269e-01 -7.142 9.18e-13 \*\*\*

agechild 2.276e+01 1.651e+04 0.001 0.99890

sexmale:classcrew 1.124e+00 8.199e-01 1.371 0.17032

sexmale:classsecond 5.593e-02 6.698e-01 0.083 0.93346

sexmale:classthird 2.885e+00 5.619e-01 5.134 2.84e-07 \*\*\*

classcrew:agechild NA NA NA NA

classsecond:agechild 2.001e+00 2.102e+04 0.000 0.99992

classthird:agechild -2.242e+01 1.651e+04 -0.001 0.99892

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 671.9616 on 13 degrees of freedom

Residual deviance: 1.7019 on 3 degrees of freedom

AIC: 70.308

Number of Fisher Scoring iterations: 21

The above result gives the summary of the stepwise logistic regression, the logistic equation is given by:

From the summary table above, provided other predictors are held constant, it can be implied that:

* The model requires 21 iterations to perform the fit.
* The rate of survival for male is significantly lower than for female. (coefficient = -4.273 and p-value =7.67e-16)
* The rate of survival for people in crew class is significantly lower than the first class. (coefficient = -1.646 and p-value =0.03955)
* The survival rate for people in the second class is significantly lower than the first class (coefficient = -1.732 and p-value =0.00317)
* The rate of survival for people in the third class is significantly lower than the first class (coefficient = -3.763 and p-value =9.18e-13)
* The rate of survival of a child is higher but not significantly higher than the proportion of adults that survived (coefficient = 0.2276 and p-value = 0.9989)
* The survival rate of male in the crew class is higher but not significantly higher than male in crew-class or female in first-class (coefficient = 1.124 and p-value = 0.17032)
* The survival rate of male in the second class is higher but not significantly higher than male in crew-class or female in first-class (coefficient = 0.05593 and p-value = 0.93346)
* The survival rate of male in the third class is significantly higher than male in crew-class or female in first-class (coefficient = 2.885 and p-value = 2.84e-07)
* There are no children in the crew class therefore, their survival rate cannot be computed compared to any sex or class.
* The survival rate of children in the second class is not significantly higher than the children survival rate in first class or adults in second-class (coefficient = 2.001 and p-value = 0.99992)
* The survival rate of children in the third class is not significantly lower than the children survival rate in first class or adults in third-class (coefficient = -0.2242 and p-value = 0.99892)

Analysis of Deviance Table

Model: binomial, link: logit

Response: surv\_prop

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 13 671.96

sex 1 434.51 12 237.45 < 2.2e-16 \*\*\*

class 3 106.25 9 131.20 < 2.2e-16 \*\*\*

age 1 18.93 8 112.27 1.358e-05 \*\*\*

sex:class 3 66.34 5 45.93 2.586e-14 \*\*\*

class:age 2 44.23 3 1.70 2.491e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

From the above ANOVA output it can be seen that the overall model is significant and the addition of the main term and interaction term is significant.

With a p-value of 0.637 when compared to the chi-square, we fail to reject the null hypothesis therefore, the model is a good fit as its p-value is >0.05

**Appendix: R Code**

Odour Score

odour\_score <- c(0,1,2,3,4,5,6,7,8,9,10)

patients\_with\_dementia <- c(24,20,26,25,27,29,27,29,30,23,18)

total <- c(75,64,86,87,93,96,92,105,115,112,130)

data\_odour = data.frame(odour\_score,patients\_with\_dementia,total)

#get proportion

data\_odour["proportion"] <- round(data\_odour$patients\_with\_dementia/data\_odour$total,3)

str(data\_odour)

data\_odour

plot(data\_odour$odour\_score,data\_odour$proportion,xlab="Odour Score",ylab = "Proportion")

# model

set.seed(600)

glm\_mod <- glm(proportion~odour\_score,family = "binomial",weights=total,data=data\_odour)

summary(glm\_mod)

source("Rfunctions.r")

round(betaCI(glm\_mod),3)

#Interpreting the Residual Deviance

pchisq(df=9,q=6.8493,lower.tail = F)

# ANOVA of Devoance Table

anova(glm\_mod,test='Chisq')

# Linear Predictor

eta <- glm\_mod$linear.predictors

eta

# Predict values with std errors

preds <- predict(glm\_mod,se.fit=T,type="response")

fit<-preds$fit

up<-preds$fit+2\*preds$se.fit

lw<-preds$fit-2\*preds$se.fit

# scatter plot

plot(data\_odour$odour\_score,data\_odour$proportion,xlab = "Odour Score",ylab="Proportion")

lines(data\_odour$odour\_score,fit,lty=1)

lines(data\_odour$odour\_score,lw,lty=2)

lines(data\_odour$odour\_score,up,lty=2)

#given p = 25% == 0.25

(log(0.25/(1-0.25))+0.62011)/-0.07599

Titanic

library(dplyr)

data\_titanic <- read.table(file.choose(),header = T)

head(data\_titanic)

str(data\_titanic)

data\_titanic$class <- as.factor(data\_titanic$class)

data\_titanic$age <- as.factor(data\_titanic$age)

data\_titanic$sex <- as.factor(data\_titanic$sex)

str(data\_titanic)

data\_titanic["surv\_prop"] <- round(data\_titanic$survive/data\_titanic$total,3)

data\_titanic

str(data\_titanic)

data\_titanic$class <- relevel(data\_titanic$class,ref="first")

summary(data\_titanic)

par(mfrow=c(1,3))

for (i in colnames(data\_titanic%>% select(class,age,sex))){

grp\_sum <- data.frame(sum = tapply(data\_titanic[["survive"]], data\_titanic[[i]], sum))

print(grp\_sum)

grp\_sum<-tibble::rownames\_to\_column(grp\_sum,i)

pct <- paste(round(100\*grp\_sum$sum/sum(grp\_sum$sum), 1),"%")

pie(grp\_sum$sum,labels=pct,col = rainbow(length(grp\_sum[[i]])),main=paste("Total Survival grouped by ",i))

legend("bottomright",legend=grp\_sum[[i]],fill=rainbow(length(grp\_sum$sum)))

}

for (i in colnames(data\_titanic%>% select(class,age,sex))){

avg\_prop = tapply(data\_titanic[["surv\_prop"]], data\_titanic[[i]], mean)

barplot(avg\_prop,col = rainbow(length(avg\_prop)),ylab="Average Proportion",xlab=i,main=paste("Average Survival grouped by ",i))

}

par(mfrow=c(1,1))

# Stepwise Logistic Regression

set.seed(600)

# minimal model (lower model)

low <- formula(~ 1)

# maximum model (upper model)

up <- formula(~ class\*age\*sex)

start <- glm(surv\_prop ~ 1, data=data\_titanic,family="binomial",weights = total)

fstep <- step(start,

direction="forward",

scope=list(lower=low,upper=up))

summary(fstep)

par(mfrow=c(2,2))

plot(fstep)

par(mfrow=c(1,1))

anova(fstep,test='Chisq')

round(1-pchisq(1.7,df=3),3)