Question 1 **[45 marks]**

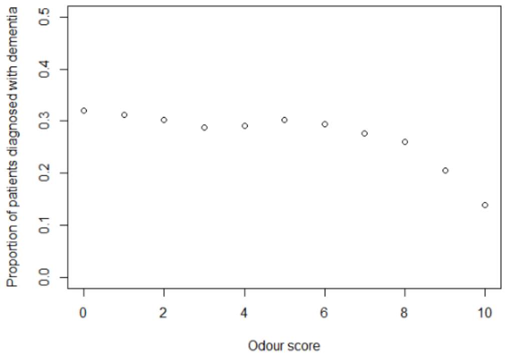

Figure 1: Proportion of patients diagnosed with dementia against the odour scores.

(a) State and fit a generalised linear model for the proportion of patients who were diagnosed with dementia against the odour score recorded. Explain all components of the model. [15 marks]

The GLM model is:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 \text{Odour score}$$

where π = population proportion of patients who diagnosed with dementia,
Odour score: is the number of correctly identified odours out of 10.

When dealing with proportions the data are generally not normally distributed. The use of **a binomial error distribution** in a GLM avoids the problem. The **logit link function** is the default for the GLM with a binomial error distribution.

Table 1: Summary table of coefficients

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6191     0.1360   -4.55  5.3e-06
odourscore   -0.0759     0.0223   -3.40  0.00068
```

From the outputs in Table 1, the model can be written as:

$$\begin{aligned}\log(\frac{\hat{p}}{1-\hat{p}}) &= \hat{\beta}_0 + \hat{\beta}_1 * \text{Odour score} \\ &= -0.619 - 0.076 * \text{Odour score}\end{aligned}$$

(b) Is the model a good fit? Refer to the relevant information from the analysis of deviance table. [10 marks]

From Table 2, the GLM model with Odour Score has a residual deviance of 6.8 on 9 df and a p-value= 0.658 which is not statistically significant. Therefore, the model is a good fit to the data.

Table 2: Analysis of Deviance for Question 1

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: dementia/total
Terms added sequentially (first to last)

            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                           10       18.4
odourscore   1     11.6         9        6.8  0.00065

> 1-pchisq(6.8,df=9)
[1] 0.658
```

(c) Give an informative interpretation of the summary table, including a practical interpretation of the coefficient of odour score.

[10 marks]

From Table 1, the fitted equation is:

$$\log(\frac{\hat{p}}{1-\hat{p}}) = -0.619 - 0.076 * \text{Odour score}$$

The p-value for Odour Score is 0.00068 which is ≈ 0. So we reject the null hypothesis H_0: $\beta_1 = 0$ and conclude that Odour Score is a significant predictor of dementia.

For an additional score in identifying odours, **the odds of being diagnosed with dementia decreases by 7.32%** ($(e^{-0.076} - 1) \times 100 = -7.32\%$). Furthermore, we can be 95% confident that there will be between 3.17% and 11.3% decrease in the odds of being diagnosed with dementia for every point increases in Odour Score (Table 3).

Table 3: 95% confident intervals for the coefficients in

```
            Estimate Std. Error  2.5 %  97.5 %
(Intercept)  -0.6191     0.1360 -0.888 -0.3550
odourscore   -0.0759     0.0223 -0.120 -0.0322

> (exp(-0.12) - 1)*100
[1] -11.3
> (exp(-0.0322) - 1)*100
[1] -3.17
```

(d) Estimate the odour score value that corresponds to 25% of patients being diagnosed with dementia. Show your calculations. [10 marks]

25% of patients being diagnosed with dementia, corresponds to $p = 0.25$. Substituting

$p = 0.25$ into the fitted model, and rearranging we have:

$$\begin{aligned}\log(\frac{\hat{p}}{1-\hat{p}}) &= -0.619 - 0.076 * \text{Odour score} \\ \log(\frac{0.25}{1-0.25}) &= \log(\frac{1}{3}) = -1.1 = -0.619 - 0.076 * \text{Odour score}\end{aligned}$$

Rearranging to solve for Odour Score:

$$OdourScore = \frac{-0.619 + 1.1}{0.076} = \frac{0.481}{0.076} = 6.33$$

Twenty-five percent of patients with an Odour Score of around 6 are predicted to suffer with dementia - refer to Figure 2. Note that the fit is *linear* in the logit scale (log(odds)).

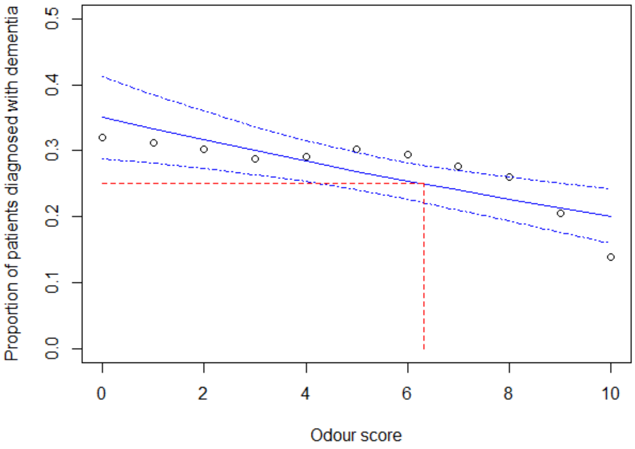

Figure 2: Proportion of patients being diagnosed with dementia together with the 95% confidence bands. Red dotted lines show 25th percentile.

Question 2: Who survived the Titanic? [50 marks]

(a) Calculate the proportion of survival in each age/sex/class combination. On the basis of examination of the data, give a brief summary of how the survival rate is affected by `class`, `age` and `sex`. [10 marks]

Sample solutions:

The proportion of survival for each age/sex/class combination is shown in Table 4. We can see that for the adults,

- 1st class passengers had the highest survival rate for both male and female (97.2% and 32.6% respectively).
- 3rd class passengers had the lowest survival rate for both male and female (46.1% and 16.2% respectively)

- female had higher survival rate than male.

For the children,

- there was no children in the crew, this will affect the regression model if we include the interaction *class*age*.
- all children in the 1st and 2nd classes survived while less than half of female children in 3rd class survived (45.2%) and only 27.1% of male children in 3rd class survived.

This suggests that there maybe an interaction *class*sex*. Other interactions might exist as well.

Table 4: Proportion of survival in each age/sex/class combination

```
adult

        female   male
crew    0.870   0.2227
first   0.972   0.3257
second  0.860   0.0833
third   0.461   0.1623
```

```
child

          female   male
crew        NA      NA
first     1.000    1.000
second    1.000    1.000
third     0.452    0.271
```

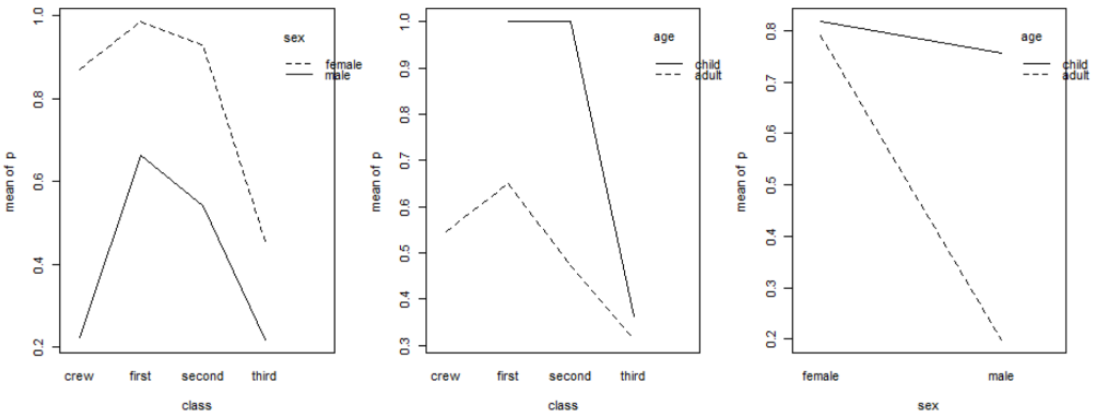

Figure 3: 2-way interaction plots for the Titanic data

(b) [40 marks]

As there was no children in the crew, we remove the interaction *class*age*. So we'll fit the GLM model of the form

$$\begin{aligned}\log(\frac{\pi}{1-\pi}) &= \beta_0 + \beta_1 * child + \beta_2 * male + \beta_3 * ClassFirst + \beta_4 * ClassSecond \\ &+\beta_5 * ClassThird + \beta_6 * child * male + \beta_7 * male * ClassFirst \\ &+\beta_8 * male * ClassSecond + \beta_9 * male * ClassThird\end{aligned}$$

using the proportion of survival as the response variable with the binomial error distribution and the logit link function.

The Analysis of Deviance Table for this model is shown in Table 5 and the summary is given in Table 6. It can be seen that both interactions age*sex and sex*class are significant (p-value $= 2.2e - 0.5$ and $2.2e - 12$ respectively).

Table 5: ANOVA table for the GLM model with 2-way interactions

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: p
Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev   Pr(>Chi)
NULL                         13        672
age        1       20        12        652    9.7e-06
sex        1      421        11        232    < 2e-16
class      3      119         8        113    < 2e-16
age:sex    1       18         7         95    2.2e-05
sex:class  3       57         4         37    2.2e-12

> 1 - pchisq(37, df=4)
[1] 1.8e-07
```

The residual deviance is 37 on 4 df resulting a p-value of $1.8e - 07$. This indicates that the model is not a good fit to the data. There are several explanations for this:

- The relationships between the predictors and the response variable may not be linear on the log odds scale, i.e. the link function was misspecified.
- Removing the interaction class*age may also lead to lack of fit of the model.

However, we'll process with the analysis in this assignment.

From Table 6, the fitted equation is:

$$\begin{aligned}\log(\frac{\pi}{1-\pi}) &= 1.90 + 0.18 * agechild - 3.15 * sexmale + 1.66 * ClassFirst + 0.05 * ClassSecond\\ &\quad -2.09 * ClassThird + 1.36 * agechild * sexmale - 1.10 * sexmale * ClassFirst\\ &\quad -0.76 * sexmale * ClassSecond + 1.56 * sexmale * ClassThird\end{aligned}$$

Only coefficients for: **sexmale, classfirst, classthird, agechild*sexmale and sexmale*classthird** are significant different from 0 (Table 6). So we convert these coefficients into the percentage of increase or decrease in odds of survival compared to the base(reference) level (Table 7).

Table 6: Summary table for the GLM model with 2-way interactions

```
Coefficients:
                     Estimate Std. Error z value   Pr(>|z|)
(Intercept)            1.8971     0.6191    3.06   0.0022
agechild               0.1803     0.3618    0.50   0.6182
sexmale               -3.1469     0.6245   -5.04   4.7e-07
classfirst             1.6642     0.8003    2.08   0.0376
classsecond            0.0497     0.6874    0.07   0.9424
classthird            -2.0894     0.6381   -3.27   0.0011
agechild:sexmale       1.3581     0.4551    2.98   0.0028
sexmale:classfirst    -1.1033     0.8199   -1.35   0.1784
sexmale:classsecond   -0.7647     0.7271   -1.05   0.2929
sexmale:classthird     1.5623     0.6562    2.38   0.0173
```

Table 7: Regression coefficients in different scales

	log odds	e^{β_i}	$(e^{\beta_i} - 1) * 100$
sexmale	-3.15	0.043	-95.7%
classfirst	1.66	5.28	428%
classthird	-2.09	0.124	- 87.6%
agechild:sexmale	1.36	3.89	289%
sexmale:classthird	1.56	4.77	377%

Interpretation:

- The intercept is the log odds of survival for **female adults in crew**.
- **Within the adults crew members**, the odds of survival for males was significantly lower than female by approximately 95.7% (p-value = 4.7e-07).
- **Compared to the adult female in the crew**, the **adult female passengers in the 1st class** had a significantly higher odds of survival by 428% (p-value = 0.0376). In contrast, the odds of survival for **adult female passengers in the 3rd class** was approximately 87.6% lower than the female crew members (p-value = 0.0011).
- For interaction $\beta_{agechild:sexmale}$, the difference in the odds of survival between adults and children is not the same for both genders (p-value = 0.0028).
- For interaction $\beta_{sexmale:classthird}$, the difference in the odds of survival between males and females is not the same for passengers in 3rd class versus the crew members (p-value = 0.0173).

It'd be interesting to see the differences in odds of survival between passengers in the first, second and third class. So you should refit the model with the reference is the first class or the third class. The residual deviance is the same in both models, so no need to include the Analysis of Deviance Table. The summary table of the re-fitted model, named Model 2, is given in Table 8.

Table 8: Summary table for Model 2 using 3rd class as the reference

```
Coefficients:
                     Estimate Std. Error z value  Pr(>|z|)
(Intercept)           -0.192      0.155   -1.24    0.2135
agechild               0.180      0.362    0.50    0.6182
sexmale               -1.585      0.202   -7.86    3.8e-15
classcrew              2.089      0.638    3.27    0.0011
classfirst             3.754      0.530    7.08    1.4e-12
classsecond            2.139      0.330    6.49    8.5e-11
agechild:sexmale       1.358      0.455    2.98    0.0028
sexmale:classcrew     -1.562      0.656   -2.38    0.0173
sexmale:classfirst    -2.666      0.567   -4.70    2.6e-06
sexmale:classsecond   -2.327      0.414   -5.62    1.9e-08
```

From Table 8, the model using the third class as the reference gives extra information, compared to Table 6.

- **For the adults in the 3rd class,** the odds of survival for male is significantly lower than female, by 80% ($(1 - exp(-1.585) * 100 = -80\%$, p-value = 3.8e-15).
- **For the adults,** the odds of survival for females passengers in the 1st and 2nd class are significantly higher than female passengers in the 3rd class by 4170% $((exp(3.754) - 1) * 100 = 4170)$and 749% $((exp(2.139) - 1) * 100 = 749)$.
- The coefficients for the interaction: **sexmale:classcrew, sexmale:classfirst** and **sexmale:classsecond** are all significantly different from 0. This suggests that the difference in odds of survival between male and female is not the same for passengers in third class versus the crew members. Furthermore, the difference in odds of survival between male and female is not the same for passengers in 1st class versus third class and 2nd class versus third class.

Note: No need to relevel sex and age as there are only 2 levels in each of them.

Appendix: R code

Question 1

```
rm(list=ls())
options(digits=3, show.signif.stars=F)
odourscore<- c(0,1,2,3,4,5,6,7,8,9,10)
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)

# Fit a GLM for the proportion
mod1 <- glm( dementia/total ~ odourscore, weights=total,family = binomial)
```

```
anova(mod1,test="Chisq")
summary.glm(mod1)
1-pchisq(6.8011,df=9)

#calculate the odds of dementia for each unit increase in odourscore
(1-exp(-0.07592))*100

# 95% confidence interval
source("Rfunctions.R")
betaCI(mod1)
(exp(-0.12) - 1)*100
(exp(-0.0322) - 1)*100

# plot observed, fitted and median values
mod1pred <- predict.glm(mod1,se.fit=T,type=’response’)
fit <- mod1pred$fit*total
plot(odourscore,dementia/total, ylim=c(0,0.5),
      ylab="Proportion of patients diagnosed with dementia",xlab="Odour score")

#Calculate approximate 95% CIs
lwr<- mod1pred$fit-2*mod1pred$se.fit
upr<-mod1pred$fit+2*mod1pred$se.fit

# add fitted line, confidence bounds
lines(odourscore,mod1pred$fit, col="blue")
lines(odourscore,lwr, lty=4, col="blue")
lines(odourscore,upr, lty=4,col="blue")

# estimate odour score for 25% of patients
#calculate 25% using log(p/(1-p))=b0+b1*odourscore
0.250/(1-0.250) # =0.3333333
log(0.250/(1-0.25)) # log value =-1.1
(1.1-0.619))/ 0.07592

lines(x=c(0,6.33),y=c(0.25, 0.25), lty=2, col = "red")
lines(x=c(6.33, 6.33),y=c(0, 0.25), lty=2, col="red")
```

Question 2:

```
########  Q2 titanic
rm(list=ls())
Titanic <- read.table("Titanic1912.txt",header=T)
Titanic$p <- Titanic$survive/Titanic$total
Titanic$class <- factor(Titanic$class)
attach(Titanic)
names(Titanic)
```

```
#Showing the proportions for different combination of categories
tapply((survive/total) , INDEX = list(class, sex, age), mean)

# plot 2-way interactions
par(mfrow=c(1,3))
with(Titanic, interaction.plot(class , sex, p))
with(Titanic, interaction.plot(class , age, p))
with(Titanic, interaction.plot(sex , age, p))

### Fit a GLM with all 2-way interactions
m1 <- glm(p ~ sex*class + class*age + sex*age,
family=binomial, weights=total)
summary(m1)
anova(m1, test='Chisq')

## remove: class*age  ##############

m2 <- glm(p ~ age*sex + sex*class,
family=binomial, weights=total)
summary(m2)
anova(m2, test='Chisq')
1 - pchisq(37, df=4)

### Refit the model with 3rd class as the baseline.

Titanic$class <-relevel(factor(Titanic$class) , ref = "third")
attach(Titanic)

m3 <- glm(p ~ age*sex + sex*class,
family=binomial, weights=total)
summary(m3)
anova(m3, test='Chisq')
#################################################################
```