Assignment 2.1

Xuening Tang, group5

4th March 2025

Exercise 1

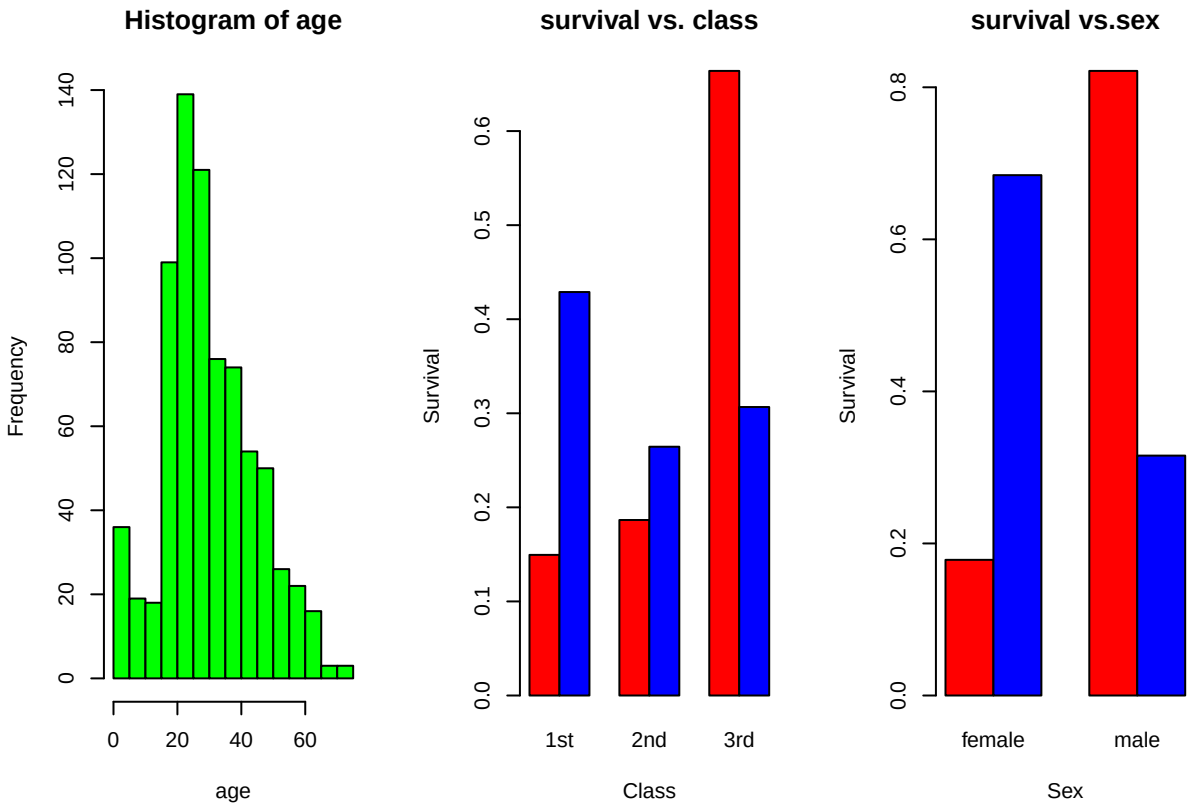
Section a We first give a summary of the data:

```
titanic <- read.delim("titanic.txt", header=TRUE)
age <- as.numeric(titanic$Age)
sex = as.factor(titanic$Sex)
class = as.factor(titanic$PClass)
survive = as.factor(titanic$Survived)
```

```
par(mfrow=c(1,3))
hist(age, col = "green")  # distribution of age

tab <- xtabs(~survive + class, data = titanic)
prop_tab <- prop.table(tab, margin = 1)
barplot(prop_tab, main="survival vs. class",
        col = c("red", "blue"),
        beside = TRUE,
        xlab = "Class", ylab = "Survival") # class vs survival

tab <- xtabs(~survive + sex, data = titanic)
prop_tab1 <- prop.table(tab, margin = 1)
barplot(prop_tab1, main="survival vs.sex",
        col = c("red", "blue"),
        beside = TRUE,
        xlab = "Sex", ylab = "Survival") # Sex vs survival
```



The first bar plot illustrates the distribution of age. Most people were around the age of 20 to 40. The second bar plot compares the survival rate in different classes (red: death; blue: survive). Most people in the 1st class survived, while most people in the 3rd class died. The plot on the right side compares the survival rate in different sexes (red: death; blue: survive). A bigger proportion of survived customers were females, while males contributed to a higher proportion of deaths.

We fitted the data with a logistic regression model, with survival as the outcome variable, age as the exploratory variable, and class and sex as the factors. The null hypotheses are:

H_a *Factor class and sex have no influence on the survival status.*

H_b *Explanatory variable age has no influence on the survival status.*

```
tiglm=glm(survive~class+age+sex,data=titanic,family=binomial)
summary(tiglm)
```

```
##
## Call:
## glm(formula = survive ~ class + age + sex, family = binomial,
##     data = titanic)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.759662   0.397567   9.457  < 2e-16 ***
## class2nd    -1.291962   0.260076  -4.968 6.78e-07 ***
## class3rd    -2.521419   0.276657  -9.114  < 2e-16 ***
## age         -0.039177   0.007616  -5.144 2.69e-07 ***
```

```
## sexmale     -2.631357   0.201505 -13.058  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  695.14  on 751  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 705.14
##
## Number of Fisher Scoring iterations: 5
```

Based on the result of the test, all three variables have **a significant effect** on the survival status, so both null hypotheses are rejected. The calculation of odd in this case is:

$$o_{ik} = \frac{P(Y_{ik}=1)}{P(Y_{ik}=0)} = e^{\mu+\alpha_i+\beta X_{ik}} \cong exp(3.76 - 1.29class2 - 2.52class3 - 0.04age - 2.63male)$$

An increase in age by 1 year would increase the odd by $e^{-0.04}$, a change from female to male will change the odd by $e^{-2.63}$, and a change from class 1 to class 2 will change the odd by $e^{-1.29}$. Since the exponential function is monotonous, all those changes will result in a lower probability of survival.

Section b We investigated the interaction effect between age and class and sex and age on the survival status. We have the following null hypotheses:

$H_0(1)$ *There is no interaction effect between age and class*

$H_0(2)$ *There is no interaction effect between age and sex*

The results are displayed below:

```
tiglm2=glm(survive~class*age,data=titanic,family=binomial)
summary(tiglm2)
```

```
##
## Call:
## glm(formula = survive ~ class * age, family = binomial, data = titanic)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.922980   0.436246   4.408 1.04e-05 ***
## class2nd     -0.744277   0.571547  -1.302 0.192843
## class3rd     -2.290072   0.540573  -4.236 2.27e-05 ***
## age          -0.035838   0.009955  -3.600 0.000318 ***
## class2nd:age -0.013209   0.015869  -0.832 0.405188
## class3rd:age  0.004642   0.015941   0.291 0.770896
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  908.75  on 750  degrees of freedom
```

```
##   (557 observations deleted due to missingness)
## AIC: 920.75
##
## Number of Fisher Scoring iterations: 4
```

```
tiglm3=glm(survive~sex*age,data=titanic,family=binomial)
summary(tiglm3)
```

```
##
## Call:
## glm(formula = survive ~ sex * age, family = binomial, data = titanic)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.30111    0.29898   1.007  0.31387
## sexmale     -0.59986    0.40805  -1.470  0.14154
## age          0.02935    0.01008   2.913  0.00358 **
## sexmale:age -0.06572    0.01369  -4.802 1.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  770.56  on 752  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 778.56
##
## Number of Fisher Scoring iterations: 4
```

The result for *age * class* **is not significant** ($p = 0.405$; $p = 0.771$). while that for *age * sex* **is significant** ($p < 0.001$). Therefore, we reject $H_0(2)$ but preserve $H_0(1)$. We incorporated the interaction effect into the model we have in section a:

```
tiglm4=glm(survive~sex*age+class,data=titanic,family=binomial)
summary(tiglm4)
```

```
##
## Call:
## glm(formula = survive ~ sex * age + class, family = binomial,
##     data = titanic)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.756563   0.437642   6.299 3.00e-10 ***
## sexmale     -0.508187   0.442515  -1.148    0.251
## age          0.002443   0.011408   0.214    0.830
## class2nd    -1.543367   0.287358  -5.371 7.83e-08 ***
## class3rd    -2.653981   0.291423  -9.107  < 2e-16 ***
## sexmale:age -0.075591   0.015009  -5.036 4.74e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  667.08  on 750  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 679.08
##
## Number of Fisher Scoring iterations: 5
```

Based on the result of this model, the main effects of sex and age are **no longer significant** when the interaction effect of *age * sex* is incorporated. We predicted the probability of survival for each combination of levels of the factors class and sex for a person aged 55:

```
new_data = expand.grid(
  age = 55,
  class = c("1st","2nd", "3rd"),
  sex = c("female","male")
)
rownames(new_data) <- c("1st&female","2nd&female","3rd&female",
                        "1st&male","2nd&male","3rd&male")
predict(tiglm4,new_data,type = "response")
```

```
## 1st&female 2nd&female 3rd&female   1st&male   2nd&male   3rd&male
## 0.94739747 0.79373500 0.55896778 0.14495247 0.03495485 0.01178897
```

The probability of survival for a female in the 1st class is the highest (0.947), while that for a male in the 3rd class is the lowest (0.012).

Section c The formula for the logistic regression is:

$$P(Y_k = 1) = \frac{1}{1 + e^{-x_k^T \theta}}, k = 1, 2 \ldots, N$$

where we want to estimate the parameter θ. We can do this by splitting our data set into two parts (90% and 10%), where the bigger part is used for training the model and estimating θ, while the smaller one is used for testing and validating our estimation. We can evaluate our model by comparing the result of our prediction with the true survival status and adjust our estimation based on the performance of our model.

Section d Another approach to study the main effect of class and sex on the survival status is conducting Chi-square tests on contingency tables:

```
# create a tab for variable sex and class
tabd <- xtabs(~class+survive, data = titanic)
tabd2 <- xtabs(~sex+survive, data = titanic)
```

```
z = chisq.test(tabd);z
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabd
## X-squared = 172.3, df = 2, p-value < 2.2e-16
```

```r
z2 = chisq.test((tabd2));z2
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  (tabd2)
## X-squared = 329.84, df = 1, p-value < 2.2e-16
```

According to the Chi-square tests, both the effect of sex and class are significant. However, this result deviates partly from what we got in section b, where we spotted an insignificant main effect of sex. This is because we didn't introduce the interaction effect in this case, unlike in section b.

Section e Even though the results we got in sections d and b are not completely consistent, the approach in section d is not necessarily wrong. A contingency table can help us test whether two variables are independent or homogeneous. An advantage is that, before conducting any statistical test or making any calculation, we may already have a first intuition of the potential relationship between the two variables. This can be done by simply comparing values in different cells. The disadvantage is that we need at least 5 items in each cell to make sure that the Chi-square result is valid. This can be a problem when we are working with data that do not have enough items for each category. Sometimes dividing the data into different categories may not be trivial either. A contingency table is also harder to interpret when we have multiple variables and want to study the interaction effects among them.

Logistic regression can incorporate a wide range of variables: binary, explanatory, and categorical (factors). A drawback is that it assumes a linear relationship between the outcome and predictors, so it may not work well for non-linear patterns.