Statistics 2: Computer Practical 3

1 Instructions

1. Reports should be handed in **as a single PDF file** using Blackboard, by noon on the due date. RMarkdown, image, word, zip files, for example, will **not** be marked.
2. You can work alone or in a group with up to two other people.
3. One person per group should hand in a report on blackboard. The names and student numbers of all group members must be on the first page.
4. Your answers should combine text and code snippets in R. It is recommended that you use RMarkdown to prepare your reports, since this is typically easier for students, but this is not mandatory.
5. You must explain what you are doing clearly to obtain full marks on each question. You can use comments (which start with #) to annotate your code. Mathematical derivations can be written using LaTeX commands in RMarkdown or on paper, with a photo then appended to the end of the PDF being submitted.
6. This practical counts for 10% of your assessment for Statistics 2.

2 Data description

This coursework focuses on analysing data about passengers on the Titanic. The Titanic was a boat that sank in the Atlanic Ocean during her maiden voyage from Southampton in England to New York in the United States. This data was extracted from the Github account of the book "Efficient Amazon Machine Learning", published by Packt: https://github.com/alexisperrier/packt-aml/blob/master/ch4/original_titanic.csv.

These data can be loaded in using the following code:

```
titanic_original = read.csv("original_titanic.csv")
```

This dataset includes information about 1309 passengers and includes information about:

- `pclass`: The class of the passengers (1 first class (most expensive), 2 second class, 3 third class (cheapest)).
- `survived`: If they survived the sinking (1 they survived, 0 they died).
- `name`: Name of the passenger.
- `sex`: Sex of the passenger (female or male)
- `age`: Age of the passenger.
- `sibsp`: Number of siblings or spouses aboard.
- `parch`: Number of parents or children aboard.
- `ticket`: The ticket number.
- `fare`: Ticket price in pounds.
- `cabin`: Cabin number.
- `embarked`: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton).
- `boat`: Boat used to escape from the Titanic.
- `body`: Body Identification Number.
- `home.dest`: Passenger final destination.

Upon observing the data, it becomes clear that it is incomplete. We assume the ages, fares and embarkation point of the passengers that were missing are at random so can remove the lines where these are NA or empty. We also remove the other columns with missing data since they are not important for our analysis (`ticket`, `cabin`, `boat`, `body` and `home.dest`).

```
titanic = titanic_original[which(!is.na(titanic_original$age) &
                                 titanic_original$embarked != "" &
                                 !is.na(titanic_original$fare)), c(1:7,9,11)]
```

This should leave you with 1043 observations.

3 Comparing exponential populations

We are first interested in if there is a difference in the ticket fares for men and women. We assume that $\mathbf{X} = (X_1, X_2.., X_n)$ and $\mathbf{Y} = (Y_1, Y_2.., Y_m)$ are independent and $X_i \overset{iid}{\sim} \text{Exp}(\lambda_1)$ and $Y_i \overset{iid}{\sim} \text{Exp}(\lambda_2)$. Let the ticket fares for men $\mathbf{x} = (x_1, ..x_n)$ and the ticket fares for women $\mathbf{y} = (y_1, ..., y_m)$ be realizations of $\mathbf{X}$ and $\mathbf{Y}$ respectively. We want to perform the following test for $\boldsymbol{\theta} = (\lambda_1, \lambda_0)$:

$$H_0 : \lambda_1 = \lambda_2 \qquad \text{vs.} \qquad H_1 : \lambda_1 \neq \lambda_2. \qquad (1)$$

The generalized likelihood ratio (GLR) in this case is

$$\Lambda(\mathbf{x}, \mathbf{y}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\lambda_1, \lambda_2; \mathbf{x}, \mathbf{y})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\lambda_1, \lambda_2; \mathbf{x}, \mathbf{y})},$$

where $\Theta = \{(\lambda_1, \lambda_2) : \lambda_1 > 0, \lambda_2 > 0\}$ and $\Theta_0 = \{(\lambda_1, \lambda_2) : \lambda_1 = \lambda_2 > 0\}$.

Question 1 [2 marks] Show that

$$\Lambda(\mathbf{x}, \mathbf{y}) = \left(\frac{\hat{\lambda}_0}{\hat{\lambda}_1}\right)^n \left(\frac{\hat{\lambda}_0}{\hat{\lambda}_2}\right)^m.$$

where $\hat{\lambda}_1 = \frac{n}{\sum_{i=1}^n x_i}$, $\hat{\lambda}_2 = \frac{m}{\sum_{i=1}^m y_i}$ and $\hat{\lambda}_0 = \frac{n+m}{\sum_{i=1}^n x_i + \sum_{i=1}^m y_i}$ and state the approximate distribution of $T = -2 \log \Lambda(\mathbf{x}, \mathbf{y})$.

You may assume that the Hessian of the log-likelihood under Θ is negative definite.

Solution The likelihood function is given by:

$$L(\lambda_1, \lambda_2; \mathbf{x}, \mathbf{y}) = \prod_{i=1}^n \lambda_1 e^{-\lambda_1 x_i} \times \prod_{i=1}^m \lambda_2 e^{-\lambda_2 y_i}$$

Taking the supremum within the constrained parameter space Θ and Θ_0, the maximized likelihoods are:

$$\sup_{\boldsymbol{\theta} \in \Theta} L(\lambda_1, \lambda_2; \mathbf{x}, \mathbf{y}) = \lambda_1^n e^{-\lambda_1 \sum_{i=1}^n x_i} \times \lambda_2^m e^{-\lambda_2 \sum_{i=1}^m y_i}$$

$$\sup_{\boldsymbol{\theta} \in \Theta_0} L(\lambda_1, \lambda_2; \mathbf{x}, \mathbf{y}) = \lambda_0^{n+m} e^{-\lambda_0 (\sum_{i=1}^n x_i + \sum_{i=1}^m y_i)}$$

Compute $\Lambda(\mathbf{x}, \mathbf{y})$:

$$\Lambda(\mathbf{x},\mathbf{y}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} L(\lambda_1,\lambda_2;\mathbf{x},\mathbf{y})}{\sup_{\boldsymbol{\theta}\in\Theta} L(\lambda_1,\lambda_2;\mathbf{x},\mathbf{y})} = \frac{\lambda_0^{n+m} e^{-\lambda_0(\sum_{i=1}^n x_i + \sum_{i=1}^m y_i)}}{\lambda_1^n e^{-\lambda_1 \sum_{i=1}^n x_i} \times \lambda_2^m e^{-\lambda_2 \sum_{i=1}^m y_i}}$$

Plugging in the expressions for λ_0, λ_1, and λ_2 yields:

$$\Lambda(\mathbf{x},\mathbf{y}) = \left(\frac{\frac{n+m}{\sum_{i=1}^n x_i + \sum_{i=1}^m y_i}}{\frac{n}{\sum_{i=1}^n x_i}} \right)^n \left(\frac{\frac{n+m}{\sum_{i=1}^n x_i + \sum_{i=1}^m y_i}}{\frac{m}{\sum_{i=1}^m y_i}} \right)^m$$

Simplifying this expression further:

$$\Lambda(\mathbf{x},\mathbf{y}) = \left(\frac{\hat{\lambda}_0}{\hat{\lambda}_1} \right)^n \left(\frac{\hat{\lambda}_0}{\hat{\lambda}_2} \right)^m$$

According to Wilks' theorem, T approximately follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters under the null and alternative hypotheses. Here, under H_0, there is 1 parameter ($\lambda_0 = \lambda_1 = \lambda_2$) and under H_1, there are 2 parameters (λ_1 and λ_2). Therefore, the approximate distribution of T is $\chi^2(1)$.

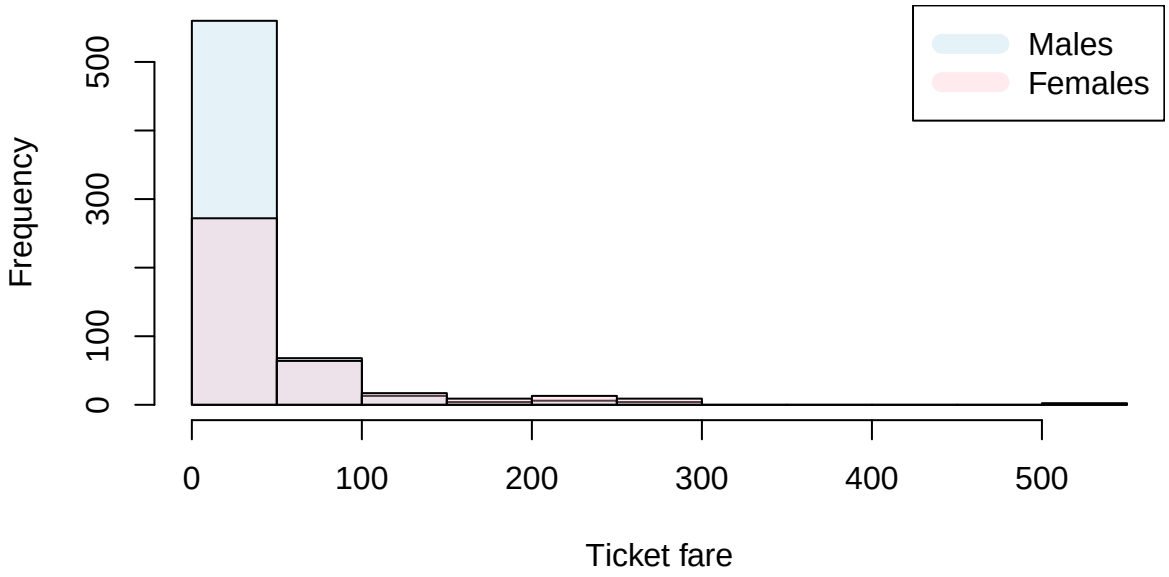
We can plot our data to gain intuition about the differences in our distributions. First we split our data into males and females.

```
fare_male <- titanic$fare[titanic$sex=='male']
fare_female <- titanic$fare[titanic$sex=='female']
```

Then we can plot histograms of the fares.

```
# Defining colours so can plot histograms over each other
c1 <- rgb(173, 216, 230, max = 255, alpha = 80, names = "lt.blue")
c2 <- rgb(255, 192, 203, max = 255, alpha = 80, names = "lt.pink")

# Plots histograms
male_hist = hist(fare_male, plot = FALSE)
female_hist = hist(fare_female, plot = FALSE)
plot(male_hist, col = c1, xlab = "Ticket fare", main = (""))
plot(female_hist, col = c2, add = TRUE)
legend("topright", legend = c("Males", "Females"), col = c(c1, c2), lty = 1, lwd = 10)
```



Question 2 [0 mark] From these figures, what result do you think the hypothesis test will return? (Note this question is worth no marks since you do the test in the next question, but to get you to think about what the data shows as well as complete the hypothesis test.)

Solution

There's a clear difference in the shape of the fare distributions between genders. There might be a potential rejection of the null hypothesis in favor of the alternative hypothesis.

Question 3 [1 mark] Perform test (1) to compare the ticket fares for males and females (`fare_male` and `fare_male`) and describe the outcome for a significance level of 0.05.

Solution

```r
# Function to calculate GLR
calculate_Lambda <- function(x, y) {
  n <- length(x)
  m <- length(y)

  lambda_1_hat <- n / sum(x)
  lambda_2_hat <- m / sum(y)
  lambda_0_hat <- (n + m) / (sum(x) + sum(y))

  Lambda <- (lambda_0_hat / lambda_1_hat)^n * (lambda_0_hat / lambda_2_hat)^m
  return(Lambda)
}
```

```r
# Function to calculate T statistic
calculate_T_statistic <- function(x, y) {
  Lambda <- calculate_Lambda(x, y)
  T <- -2 * log(Lambda)
  return(T)
}

# Calculating GLR
GLR <- calculate_Lambda(fare_male, fare_female)
cat("GLR:", GLR, "\n")
```

```
## GLR: 6.520773e-18
```

```r
# Calculating the T statistic
T_statistic <- calculate_T_statistic(fare_male, fare_female)
cat("T statistic:", T_statistic, "\n")
```

```
## T statistic: 79.14308
```

```r
# Calculating the p-value under a chi-square distribution with 2 degrees of freedom
p_value <- pchisq(T_statistic, df = 2, lower.tail = FALSE)
cat("p-value:", p_value, "\n")
```

```
## p-value: 6.520773e-18
```

```r
# Checking significance level
if (p_value < 0.05) {
  cat("The test is statistically significant at the 0.05 level. Reject the null hypothesis.\n")
} else {
  cat("The test is not statistically significant at the 0.05 level. Fail to reject the null hypothesis.
}
```

```
## The test is statistically significant at the 0.05 level. Reject the null hypothesis.
```

---

4 Testing multinomial distributions

We assume that the number of passengers with each ticket `class` come from $Y \sim$Multinomial($\mathbf{p}$) where $\mathbf{p} = (p_1, p_2, p_3)$. Here $\mathbf{p}$ is a vector of non negative probabilities that sums to 1. Alice suggests that the probability of a passenger being in each class is uniformly distributed.

Formally, we would like to test

$$H_0 : \mathbf{p} = \mathbf{p}_0 = (1/3, 1/3, 1/3) \qquad \text{vs.} \qquad H_1 : \mathbf{p} \neq \mathbf{p}_0. \qquad (2)$$

.

We can tabulate the number of passengers in each category using the following code:

```r
tab = table(titanic$pclass)
tab
```

```
##
##   1   2   3
## 282 261 500
```

Question 4. [2 marks] Use the Pearson's chi-squared test statistic to perform hypothesis test (2) for a significance level of 0.05.

Solution

```r
# Extracting observed
observed <- table(titanic$pclass)

# Assuming null hypothesis probabilities
p_0 <- c(1/3, 1/3, 1/3)

# Total number of observations
n <- sum(observed)

# Calculating expected frequencies under null hypothesis
expected <- p_0 * n

# Calculating the chi-squared test statistic
chi_squared <- sum((observed - expected)^2 / expected)

# Degrees of freedom for the chi-squared test
degrees_of_freedom <- length(observed) - 1

# Calculating p-value
p_value <- 1 - pchisq(chi_squared, df = degrees_of_freedom)


# Conducting hypothesis test
if (p_value < 0.05) {
  cat("Reject H0: The distributions are not equal (p-value =", p_value, ")\n")
} else {
  cat("Cannot reject H0: The distributions are equal (p-value =", p_value, ")\n")
}
```

```
## Reject H0: The distributions are not equal (p-value = 0 )
```

```r
# Output the chi-squared test statistic and p-value
cat("Chi-squared test statistic:", chi_squared, "\n")
```

```
## Chi-squared test statistic: 100.7536
```

```r
cat("Degrees of freedom:", degrees_of_freedom, "\n")
```

```
## Degrees of freedom: 2
```

```r
cat("P-value:", p_value, "\n")
```

```
## P-value: 0
```

Question 5 [1 mark] Bob suggests that a passenger is twice as likely to be in third class than in first class, and also twice as likely to be in third class than in second class. Evaluate this statement formally using hypothesis testing. You should define your hypotheses, calculate your Pearson test statistic and evaluate your p value.

Solution

```r
# Counting the number of passengers in each class
first_class <- sum(titanic$pclass == 1)
second_class <- sum(titanic$pclass == 2)
third_class <- sum(titanic$pclass == 3)

# Calculating observed frequencies
observed <- c(first_class, second_class, third_class)

# Calculating total number of passengers
total_passengers <- sum(observed)

# Calculating expected frequencies based on Bob's suggestion
expected <- c(total_passengers * 1/4, total_passengers * 1/4, total_passengers * 1/2)  #

# Calculating chi-square test statistic
chi_squared_statistic <- sum((observed - expected)^2 / expected)
cat("Chi-squared statistic:", chi_squared_statistic, "\n")
```

```
## Chi-squared statistic: 2.618408
```

```r
# Degrees of freedom
df <- length(observed) - 1
cat("Degrees of freedom:", df, "\n")
```

```
## Degrees of freedom: 2
```

```r
# Calculating p-value
p_value <- 1 - pchisq(chi_squared_statistic, df)
cat("p-value:", p_value, "\n")
```

```
## p-value: 0.2700349
```

```r
# Checking significance level
if (p_value < 0.05) {
  cat("The test is statistically significant at the 0.05 level. Reject the null hypothesis.\n")
} else {
  cat("The test is not statistically significant at the 0.05 level. Fail to reject the null hypothesis.
}
```

```
## The test is not statistically significant at the 0.05 level. Fail to reject the null hypothesis.
```

---

5 Logistic regression

Next we are interested in working out which of our outcome variables are significant in determining if a person survived the sinking of the Titanic. Since `survived` is a binary outcome (0 or 1) we consider a logistic model for this. Here $Y_1, \ldots, Y_n$ are independant random variables from

$$Y_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\sigma(\theta^T x_i)), \qquad i \in \{1, \ldots, n\},$$

where $x_1, \ldots, x_n$ are d-dimensional real (non random) vectors of explanatory variables such as the `age`, `sex` and `fare` of the passengers and σ is the standard logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

We define the $n \times d$ matrix $X = (x_{ij})$.

Question 6. [1 mark] Show that the log-likelihood function is

$$\ell(\theta; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i)).$$

Solution Given:

$$Y_i \sim \text{Bernoulli}(\sigma(\theta^T x_i))$$

The probability mass function (PMF) for a Bernoulli random variable is:

$$P(Y_i = y_i) = \begin{cases} \sigma(\theta^T x_i) & \text{if } y_i = 1 \\ 1 - \sigma(\theta^T x_i) & \text{if } y_i = 0 \end{cases}$$

The likelihood function for a single observation is the probability mass function:

$$L(\theta; y_i, x_i) = \sigma(\theta^T x_i)^{y_i} \cdot (1 - \sigma(\theta^T x_i))^{1-y_i}$$

Taking the logarithm (log-likelihood) of this expression:

$$\ell(\theta; y_i, x_i) = y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i))$$

Sum up the log-likelihood contributions from each observation:

$$\ell(\theta; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i))$$

This expression represents the log-likelihood function for the logistic regression model given the dataset $\mathbf{y}$ and the corresponding explanatory variables $X = (x_1, x_2, \ldots, x_n)$.

Question 7. [1 mark] Show that each component of the score is

$$\frac{\partial \ell(\theta; \mathbf{y})}{\partial \theta_j} = \sum_{i=1}^{n} [y_i - \sigma(\theta^T x_i)] x_{ij}, \qquad j \in \{1, \ldots, d\}. \quad (A)$$

I suggest that you use the fact that $\frac{d\sigma(z)}{dz} = \sigma(z)[1 - \sigma(z)]$.

Solution Given:

$$\ell(\theta; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i))$$

Differentiate $\ell(\theta; \mathbf{y})$ with respect to θ_j:

$$\frac{\partial \ell(\theta; \mathbf{y})}{\partial \theta_j} = \sum_{i=1}^{n} \left[y_i \frac{1}{\sigma(\theta^T x_i)} \frac{\partial \sigma(\theta^T x_i)}{\partial \theta_j} + (1 - y_i) \frac{1}{1 - \sigma(\theta^T x_i)} \frac{\partial (1 - \sigma(\theta^T x_i))}{\partial \theta_j} \right]$$

Using the chain rule of differentiation and the fact that $\frac{d\sigma(z)}{dz} = \sigma(z)[1 - \sigma(z)]$, proceed:

$$\frac{\partial \ell(\theta;\mathbf{y})}{\partial \theta_j} = \sum_{i=1}^{n}\left[y_i \frac{1}{\sigma(\theta^T x_i)}\sigma(\theta^T x_i)[1-\sigma(\theta^T x_i)]x_{ij} - (1-y_i)\frac{1}{1-\sigma(\theta^T x_i)}\sigma(\theta^T x_i)[1-\sigma(\theta^T x_i)]x_{ij}\right]$$

Simplify the expression:

$$\frac{\partial \ell(\theta;\mathbf{y})}{\partial \theta_j} = \sum_{i=1}^{n}\left[y_i(1-\sigma(\theta^T x_i)) - (1-y_i)\sigma(\theta^T x_i)\right]x_{ij}$$

Notice that $y_i(1-\sigma(\theta^T x_i)) - (1-y_i)\sigma(\theta^T x_i) = y_i - \sigma(\theta^T x_i)$, therefore:

$$\frac{\partial \ell(\theta;\mathbf{y})}{\partial \theta_j} = \sum_{i=1}^{n}[y_i - \sigma(\theta^T x_i)]x_{ij}$$

This matches the expression given in equation (A), confirming that each component of the score is indeed $\frac{\partial \ell(\theta;\mathbf{y})}{\partial \theta_j} = \sum_{i=1}^{n}[y_i - \sigma(\theta^T x_i)]x_{ij}$.

From Equation A, the score can be written as

$$\nabla \ell(\theta;\mathbf{y}) = X^T[\mathbf{y} - \mathbf{p}(\theta)],$$

where $\mathbf{p}(\theta)$ is the vector $(p_1(\theta), \ldots, p_n(\theta)$ and $p_i(\theta) = \sigma(\theta^T x_i)$.

6 Hypothesis testing in logistic regression

If an explanatory variable has no effect on the probability of the response variable then we expect the corresponding coefficient to be equal to 0. This can be examined more formally using hypothesis testing.

Assume that we consider the logistic model described in Section 5 and we want to test if `age` is a significant variable. It is useful to add a column of 1s to X, so that there is an "intercept" term in the model. Mathematically, the value of θ_0 determines the probability when the explanatory variables are all 0. Therefore, our vector of parameters becomes $\boldsymbol{\theta} = (\theta_0, \theta_{\text{sex}}, \theta_{\text{age}}, \theta_{\text{fare}})$ and our hypothesis test is

$$H_0 : \theta_{\text{age}} = 0 \qquad \text{vs.} \qquad H_1 : \theta_{\text{age}} \neq 0. \qquad (3)$$

We can consider the generalised likelihood ratio statistic for this test,

$$\Lambda_n = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta};\boldsymbol{y})}{\sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta};\boldsymbol{y})} = \frac{L(\hat{\boldsymbol{\theta}}_0;\boldsymbol{y})}{L(\hat{\boldsymbol{\theta}}_{MLE};\boldsymbol{y})}$$

where $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimator under the null hypothesis, $\hat{\boldsymbol{\theta}}_{MLE}$ is the maximum likelihood estimator for the full model (which we derived in Section 5) and $\Theta = \{(\theta_0, \theta_{\text{sex}}, \theta_{\text{age}}, \theta_{\text{fare}}) : \theta_0 \in \mathbb{R}, \theta_{\text{sex}} \in \mathbb{R}, \theta_{\text{age}} \in \mathbb{R}, \theta_{\text{fare}} \in \mathbb{R}\}$ and $\Theta_0 = \{(\theta_0, \theta_{\text{sex}}, \theta_{\text{age}}, \theta_{\text{fare}}) : \theta_0 \in \mathbb{R}, \theta_{\text{sex}} \in \mathbb{R}, \theta_{\text{age}} \in \{0\}, \theta_{\text{fare}} \in \mathbb{R}\}$.

Then,

$$-2\log\Lambda_n = -2\{l(\hat{\boldsymbol{\theta}}_0;\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}_{MLE};\boldsymbol{y})\}$$

has a $\mathcal{X}^2_r$ distribution **under the null hypothesis** (notice that r is the number of restrictions under the null hypothesis).

Here we extract the variables we are interested in for the analysis, encode `sex` as a numeric value (0 for men and 1 for women) and add our intercept.

```
survived = titanic$survived
sex = ifelse(titanic$sex == "male", 0, 1)
age = titanic$age
fare = titanic$fare
intercept = rep(1, length(survived))
data = data.frame(intercept, sex, age, fare, survived)
```

Question 8 [2 marks] Use the generalised likelihood ratio test to decide whether the ticket `age` is statistically significant for surviving the titanic when `sex`, `age` and `fare` are considered for a significance level of 0.05.

To do this start by forming the matrices that correspond to the restricted model under the null hypothesis (`X_rest`) and to the full model (`X_full`). In this case `X_rest` is formed by removing the variable `age` because this is the one we want to test.

```
X_full <- as.matrix(data[c('intercept', 'sex', 'age', 'fare')])
X_rest <- as.matrix(data[c('intercept', 'sex', 'fare')])
Y <- data[, 'survived']
```

You may wish to use the following functions for your calculation.

```
sigma <- function(v) {
  1 / (1 + exp(-v))
}

ell <- function(theta, X, y) {
  p <- as.vector(sigma(X%*%theta))
  sum(y*log(p) + (1-y)*log(1-p))
}

score <- function(theta, X, y) {
  p <- as.vector(sigma(X%*%theta))
  as.vector(t(X)%*%(y-p))
}

maximize.ell <- function(ell, score, X, y, theta0) {
  optim.out <- optim(theta0, fn=ell, gr=score, X=X, y=y, method="BFGS",
                     control=list(fnscale=-1, maxit=1000, reltol=1e-16))
return(list(theta=optim.out$par, value=optim.out$value))
}
```

Solution

```
# Extract relevant columns and create the data frame
survived <- titanic$survived
sex <- ifelse(titanic$sex == "male", 0, 1)
age <- titanic$age
fare <- titanic$fare
intercept <- rep(1, length(survived))

# Create a data frame
data <- data.frame(intercept, sex, age, fare, survived)

# Define the full and restricted matrices
X_full <- as.matrix(data[c('intercept', 'sex', 'age', 'fare')])
X_rest <- as.matrix(data[c('intercept', 'sex', 'fare')])
```

```r
Y <- data[, 'survived']

sigma <- function(v) {
  1 / (1 + exp(-v))
}

ell <- function(theta, X, y) {
  p <- as.vector(sigma(X %*% theta))
  sum(y * log(p) + (1 - y) * log(1 - p))
}

score <- function(theta, X, y) {
  p <- as.vector(sigma(X %*% theta))
  as.vector(t(X) %*% (y - p))
}

maximize.ell <- function(ell, score, X, y, theta0) {
  optim.out <- optim(theta0, fn = ell, gr = score, X = X, y = y, method = "BFGS",
                     control = list(fnscale = -1, maxit = 1000, reltol = 1e-16))
  return(list(theta = optim.out$par, value = optim.out$value))
}

# Fit the models
theta_rest <- maximize.ell(ell, score, X_rest, Y, theta0 = rep(0, ncol(X_rest)))
theta_full <- maximize.ell(ell, score, X_full, Y, theta0 = rep(0, ncol(X_full)))

# Calculate log-likelihoods
l_rest <- ell(theta_rest$theta, X_rest, Y)
l_full <- ell(theta_full$theta, X_full, Y)

# Calculate the likelihood ratio test statistic
LR_statistic <- -2 * (l_rest - l_full)

# Obtain the degrees of freedom
df <- ncol(X_full) - ncol(X_rest)

# Calculate the p-value
p_value <- pchisq(LR_statistic, df = df, lower.tail = FALSE)

# Decide based on the significance level (0.05)
if (p_value <= 0.05) {
  cat("Age is statistically significant for survival with a p-value of", p_value, "\n")
} else {
  cat("Age is not statistically significant for survival with a p-value of", p_value, "\n")
}
```

```
## Age is statistically significant for survival with a p-value of 0.03652914
```

---

7 Epilogue

There are some functions in R that can be used to calculate the p-values for questions 4, 5 and 8. These can only be used to check your answers. This section is worth no marks and does not need to be completed.

7.1 Pearson's Chi-squared Test for Count Data

We can obtain the p-values for a Pearson's Chi-squared text using the function chisq.test where counts is a vector of counts for each category and p is a vector of the proposed probabilities. The length of the vectors counts and probabilities must be the same. Use `?chisq.test` for more information.

```
p = chisq.test(counts, p)
```

7.2 Likelihood Ratio Test of Nested Models

We can obtain the estimated values for the parameters in logistic regression using the glm function. One can then use the lrtest (need to load lmtest package) to perform hypothesis testing for nested models. For example, if we have two explanatory variables x1, x2 and we want to test if x1 is significant for the response y, we can test this in R as follows:

```
library(lmtest)
model_full = glm(y ~ x1 + x2, data = data, family=binomial(link='logit'))
model_restricted = glm(y ~ x1, data = data, family=binomial(link='logit'))

lrtest(model_full, model_restricted)
```