HW1_cq2203

Chang Qu (cq2203)

2018/9/15

Contents

Problem 1

Part 1: Importing Data into R

i. Import the titanic dataset into RStudio using read.table(). Use the argumentas.is= TRUE. The dataset should be stored in a data frame called titanic.

```
titanic <- read.table("/Users/courtneyqu/Desktop/GR5206/Titanic.txt", header = TRUE, as.is = TRUE)
```

ii. How many rows and columns doestitanichave? (If there are not 891 rows and 12 columns something is wrong. Check part (i) to see what could have gone wrong.)

```
#check the number of rows and columns of this dataset
nrow(titanic)
```

```
## [1] 891
```

```
ncol(titanic)
```

```
## [1] 12
```

iii. Create a new variable in the data frame called Survived.Word. It should read either “survived” or “died” indicating whether the passenger survived or died. This variableshould be of type ‘character’.

```
#create new variable
n <- length(titanic$PassengerId)
#creat a container for the new variable
Survive.Word <- rep(0, n)
#assign value to the new variable
for(i in 1:n){
  if(titanic$Survived[i] == 0){
    Survive.Word[i] <- "died"
  }
  else{
    Survive.Word[i] <- "survived"
  }
}

#check the type of new variable
typeof(Survive.Word)
```

```
## [1] "character"
```

```
#adding the new variable to data frame
titanic <- cbind(titanic, Survive.Word)
```

Part 2: Exploring the Data in R

i. Use the apply() function to calculate the mean of the variables Survived, Age, and Fare.

```
#calculate the mean of the variable survived, age, and fare
apply(titanic[,c(2,6,10)], 2, mean)
```

```
##   Survived         Age       Fare
##  0.3838384          NA 32.2042080
```

From the result, we can see that the mean of age is NA. Because there are values of NA in the varivale Age, which will return a value of NA when calculating the mean.

ii. Compute the proportion of female passengers who survived the titanic disaster.

```
round(mean(titanic$Survived[titanic$Sex == "female"]), 2)
```

```
## [1] 0.74
```

Therefore, the proportion of female passenger who survived the Titanic is 0.74.

iii. Of the survivors, compute the proportion of female passengers. Round your answer to 2 decimals.

```
#split the passengers into survivors and died
group <- split(titanic$Sex, factor(titanic$Survive.Word))

#create table of survivors according to gender
table_n <- table(group$survived)

#calculate the probabilty
round(table_n/length(group$survived),2)
```

```
##
## female   male
##   0.68   0.32
```

Therefore, of the survivors, the proportions of female passengers is 0.68.

iv. Create an empty numeric vector of length three called Pclass.Survival, which elements are the survival rates of the three classes by using a loop.

```
classes <- sort(unique(titanic$Pclass))
Pclass.Survival <- vector("numeric", length = 3)
names(Pclass.Survival) <- classes

#split passengers accroding to their class with levels of survived and died
class.group <- split(titanic$Survive.Word, factor(titanic$Pclass))

for(i in 1:3){
  class_table <- table(class.group[i])
  new_table <- round(class_table / length(unlist(class.group[i])), 2)
  Pclass.Survival[i] <- new_table["survived"]
}
Pclass.Survival
```

```
##    1    2    3
## 0.63 0.47 0.24
```

From the result above, we can see that the survival rate for class 1 is 0.63, the survival rate for class 2 is 0.47, the survival rate for class 3 is 0.24.

v. Now create a Pclass.Survival2 vector that should equal the Pclass.Survival vector from the previous question. But using vectorized operation.

```r
#create a function for conditional probability
survive.prob <- function(list_name){
  table_1 <- table(list_name)
  table_2 <- round(table_1 / sum(table_1),2)
  return(table_2["survived"])
}

Pclass.Survival2 <- tapply(titanic$Survive.Word, factor(titanic$Pclass), survive.prob)
Pclass.Survival2
```

```
##    1    2    3
## 0.63 0.47 0.24
```

vi. Does there appear to be a relationship between survival rate and class?

```r
#fit a linear regression model between the Pclass and the survival rate
class.var <- c(1, 2, 3)
rate.var <- c(0.63, 0.47, 0.24)
model <- lm(rate.var~class.var)
summary(model)
```

```
##
## Call:
## lm(formula = rate.var ~ class.var)
##
## Residuals:
##        1        2        3
## -0.01167  0.02333 -0.01167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.83667    0.04365   19.17   0.0332 *
## class.var   -0.19500    0.02021   -9.65   0.0657 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02858 on 1 degrees of freedom
## Multiple R-squared:  0.9894, Adjusted R-squared:  0.9788
## F-statistic: 93.12 on 1 and 1 DF,  p-value: 0.06574
```

In this case, we want to do fit a linear regression model between the variable class and the survival rate, then do a hypothesis test on β_1, which is:

$H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$

As we can see from the result above, $P - value = 0.0657$. So the probability of observing a value of -0.195 or more extreme as the estimator for β_1 is no more than 6.57%, which is greater than the significant level of 5%. So we cannot conclude that there is a relationship between the survival rate and the class.