HW4_IS457_8

Part 1: Sampling and Point Estimation

The following problems will use the diamonds dataset which

containing the prices and other attributes of almost 54,000 diamonds.

Load the data by running the following code

```
#install.packages("ggplot2") # Delete or comment this line if you have
installed this package
library(ggplot2)
data(diamonds)
```

We would like to split our dataset into two groups and make some analysis about the

price of diamonds of each group.

1. (1 pt)

Use the sample function to generate a vector named “group” of 1s and 2s that has the same

length as the diamonds dataset.

IMPORTANT: Make sure to run the following seed function before you run your sample

function. Run them back to back each time you want to run the sample function to ensure

the same seed is used every time.

```
set.seed(666)
```

Your answer here

```
group = sample(1:2, size= 53940, replace = TRUE)
```

2. (1 pt)

Use this vector to split the 'price' variable into two vectors, Price1 and Price2.

(Hint: if elements in vector 'group' is 1, put the correspont element in 'Price' into 'Price1', ect.)

Check: If you did this properly, you will have 26753 elements in Price1 and 27187 elements

in Price2. You can print the length here, please don't print out the vector(it's too long!).

Your answer here

```
       Price1= c(diamonds$price[group == 1])
       Price2= c(diamonds$price[group == 2])
       length(Price1)
```

```
## [1] 26753
```

```
       length(Price2)
```

```
## [1] 27187
```

3. (3 pts)

Create two 95% confidence intervals for mean of Price1 and Price2.

(you can use the following formula for a confidence interval: mean ± 1.96 x SE).

Compare the confidence interval of Price1 and Price2 –

do they seem to agree or disagree?

Your answer here

```
    #Price1
     SE_price1 = sd(Price1,na.rm = TRUE)/sqrt(length(Price1[!is.na(Price1)]))
       mean(Price1)- (1.96 * SE_price1 )
```

```
## [1] 3882.945
```

```
       mean(Price1) + (1.96 * SE_price1 )
```

```
## [1] 3978.763

 #95% confidence interval for mean of price1 is (3883,3979)

#Price2
       SE_price2 = sd(Price2,na.rm=TRUE)/sqrt(length(Price2[!is.na(Price2)]))
       mean(Price2) - (1.96 * SE_price2)

## [1] 3887.392

       mean(Price2) + (1.96 * SE_price2)

## [1] 3982.037

  #95% confidence interval for mean of price2 is (3887,3982)
  #They agree
```

4. (2 pts)

Write a funciton to calculate the 95% confidence intervals of

any input vector x.

Your answer here

```
CIfunction=function(x){
    SE = sd(x,na.rm = TRUE)/sqrt(length(x[!is.na(x)]))
              quant = qt(0.975,(length(x))-1)
             lower_bound = mean(x) - (quant * SE)
             upper_bound= mean(x) + (quant * SE)

  ## return your confidence interval as below
  return(c(lower_bound,upper_bound))
}
```

5. (1 pt)

Draw 5000 observations from a standard normal distribution.

Use the function you wrote in Q4 to calculte

the 95% confidence intervals of these samples.

Your answer here

```
set.seed(666)
random_observations = rnorm(5000)
CIfunction(random_observations)

## [1] -0.01849028  0.03702744
```

Part 2: Life Cycle of Data Science

These questions will practice working with the concept of the

Life Cycle of Data Science.

“Regression” refers to the simple linear regression equation:

y = b0 + b1*x which is the model we have seen in class

6. (2 pts)

Choose an appropriate Life Cycle of Data Science from the lecture notes,

or describe one of your own, and explain how and where a linear regression

model fits into the Life Cycle you have chosen.

Your answer here

#acquire–> clean–>use/reuse–>publish–>preserve/destroy # the linear regression model will fit into the use/reuse phase which is the point at which we analyze,visualize #the data. We apply models to the data in order to observe relationships between variables, trends and # make predictions.

7. (2 pt)

What data and/or code are associated with this linear regression model? Where do they

fit in the Life Cycle of Data Science?

Your answer here

The data associated with the linear regression models are several observations from 2 variables which are to be analyzed. #The datasets are gotten during the acquire phase of the data science Life cycle # code such as below #lm(weight ~ height, data = family)

Part 3: Boston Housing

This part we will use the Boston Housing data set.

Load data by running code below

```
housing <- read.table(url("https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data"),sep="")
names(housing) <- c("CRIM","ZN","INDUS","CHAS","NOX","RM","AGE","DIS","RAD","TAX","PTRATIO","B","LSTAT","MEDV")
```

We will focus on two variables:

MEDV: median value of owner-occupied homes in $1000's

CRIM: per capita crime rate by town

8. (2 pt)

Find the correlation between crime rate and house value, and explain what does

it mean in the context of data.

Your answer here

```
cor(housing$CRIM,housing$MEDV)
```

```
## [1] -0.3883046
```

#a negative value of -0.3883046 indicates a weak negative correlation between the two variables # meaning an increase in the per capita crime rate causes a decrease in median value of homes but crime rate # is only an about 39% accurate predictor of house value

9. (2 pts)

Suppose we also want to consider distance to city centres(DIS) as another

factor that may affect house value.

Fit a linear regression model of Boston Housing data

using crime rate and distance(DIS) to predict house value, using lm().

Your answer here

```
   lm(MEDV ~ CRIM + DIS, data = housing)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + DIS, data = housing)
##
## Coefficients:
## (Intercept)         CRIM          DIS
##     21.8722      -0.3666       0.5231
```

we can then say MEDV = 21.8722 + (-0.3666CRIM) + 0.5231DIS #MEDV = 21.8722 - 0.3666CRIM + 0.5231DIS

10. (2 pt)

Use information from summary() of the model to explain how crime rate and

distance to city centres will affect the house value. How does this compare to the

correlation data from question 8?

Your answer here

```
summary(lm(MEDV ~ CRIM + DIS, data = housing))
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + DIS, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.723  -5.380  -2.035   2.160  30.901
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.87224    0.89342  24.482  < 2e-16 ***
## CRIM        -0.36657    0.04715  -7.775 4.28e-14 ***
## DIS          0.52310    0.19258   2.716  0.00683 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.431 on 503 degrees of freedom
## Multiple R-squared:  0.1631, Adjusted R-squared:  0.1597
## F-statistic:    49 on 2 and 503 DF,  p-value: < 2.2e-16
```

the model p-value and the p-values of the variables show that the linear model is statistically significant. # Crime Rate like in question 8 has a weak negative coefficient but DIS has a moderately positive coefficient #meaning that the predicted median value of homes decreases based on increase in crime rate and increases based on increase in DIS

11. (2 pts)

Describe the LifeCycle of Data for Part 3 (Question 8-10) of this homework.

Your answer here

acquire–> clean–>use/reuse–>publish–>preserve/destroy

we gathered/collected our data(acquire phase) and then we built a linear regression model to predict a variable based on 2 other predictor variables(use/reuse),building our model is part of analyzing our data. We also tried to ascertain the correlation between variables.