**STAT 40001/50001 Statistical Computing Fall 2024**

**Test 2- Answer key**

**Name:**

**PUID: 012345678**

*This exam consists of 4 questions. Please provide the R codes that have been used to perform all the calculations and graphics along with the interpretation of the output. It is important that you distinguish between the R code and the result description. You may use different fonts or different colors.*

**Q.N. 1) (30 points)** The data provided in the link below describes the fuel consumption (in mpg) based on the weight (in lbs) for different brands of cars.

<http://www.stat.cmu.edu/~cshalizi/mreg/15/hw/04/auto-mpg.csv>

1. Import the data in R and display the information using a scatter plot.

**Important: You will use first 250 observations of the dataset if your PUID has last digit 5 or higher otherwise you will use the last 250 observations to answer the following questions.**

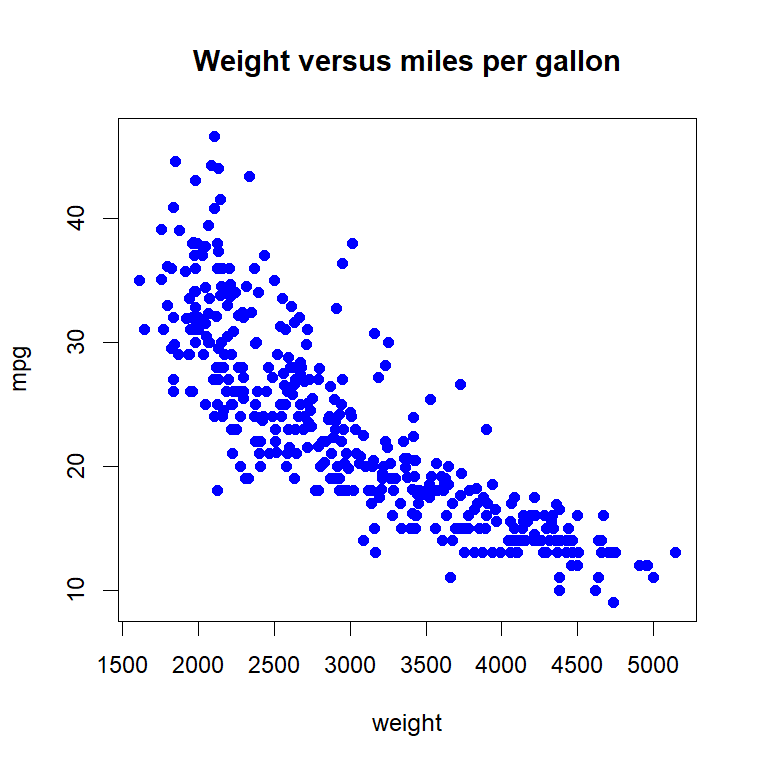
1. Select 250 observations based on the above instruction and print the first five observations of your dataset.
2. Fit a simple linear regression model and state *the equation of the model*.
3. Please provide the interpretation of the parameter () to determine the relationship between weight and fuel consumption.
4. Determine the coefficient of determination of the model and provide its interpretation.
5. Use the model to predict the mpg if the car is 2100 lbs. Please provide a 90% confidence interval for your predicted value.

*Solution: a) We used R code below to import the data*

*> Q1=read.csv("http://www.stat.cmu.edu/~cshalizi/mreg/15/hw/04/auto-mpg.csv")*

*> attach(Q1)*

*> plot(weight, mpg, col="blue", pch=16, main="Weight versus miles per gallon")*

**

*b) Since the PUID is an even number first 250 observations will be used*

*> Q1new=Q1[1:250,]*

*> dim(Q1new)*

*[1] 250 3*

*> head(Q1new, 5)*

*car.name mpg weight*

*1 chevrolet chevelle malibu 18 3504*

*2 buick skylark 320 15 3693*

*3 plymouth satellite 18 3436*

*4 amc rebel sst 16 3433*

*5 ford torino 17 3449*

*c) Based on the R output below the fitted model is mpg=39.983609-0.006206\*weight*

*> model=lm(mpg~weight, data=Q1new)*

*> model*

*Call:*

*lm(formula = mpg ~ weight, data = Q1new)*

*Coefficients:*

*(Intercept) weight*

*39.983609 -**0.006206*

*d.) Based on the results mpg will be decreased at the rate of 0.006206 per lbs increase in the weight. This is equivalent to a decrease in 12.412 (=0.006206/5\*10000) per ton increase in the weight.*

e) *Based on the R output below the coefficient of determination is 0.7885. This means that 78.85 % variation in the mpg can be explained using the weight of the vehicle.*

*> summary(model)$r.sq*

*[1] 0.7885352*

f) *Based on the R output below it has been predicted that the mpg for a 2100 lbs vehicle is 26.95 with 90% confidence interval (26.48, 27.42).*

*> predict(model,data.frame(weight=2100),interval="conf", level=0.9)*

*fit lwr upr*

*1 26.95165 26.48383 27.41947*

**Q.N. 2)** (**25 points**) A small-scale clinical trial is conducted to study the effect of drug in reduction of excess body weight. Age and gender (0- Female, 1-Male) were recorded at the baseline. The percent excess body weight loss was recorded after 3 months into the study. The data are provided below

|  |  |  |
| --- | --- | --- |
| age | gender | EWL |
| 49 | 0 | 24.2 |
| 54 | 1 | 25.4 |
| 37 | 0 | 20.1 |
| 43 | 0 | 21 |
| 57 | 1 | 27.7 |
| 48 | 1 | 16.6 |
| 34 | 0 | 15.9 |
| 51 | 0 | 27.4 |
| 54 | 0 | 27.8 |
| 45 | 0 | 25.3 |
| 36 | 1 | 12.7 |
| 57 | 1 | 25 |
| 44 | 1 | 18.4 |
| 56 | 1 | 28.2 |
| 44 | 1 | 19.3 |
| 47 | 1 | 22.5 |
| 44 | 0 | 26.7 |
| 52 | 0 | 25.4 |
| 51 | 1 | 21.9 |
| 44 | 0 | 26.6 |
| 53 | 0 | 26.8 |
| 55 | 1 | 27.4 |
| 30 | 0 | 23.1 |
| 47 | 1 | 19.8 |
| 26 | 1 | 14.1 |
| 56 | 0 | 29.6 |
| 28 | 0 | 17.8 |
| 34 | 1 | 19.8 |
| 43 | 1 | 20.6 |
| 55 | 1 | 26.8 |
| 52 | 0 | 24.7 |
| 54 | 0 | 26.7 |

1. Fit a multiple linear regression model reflecting the effect of gender. Be sure to write the equation of the model.
2. Please be sure to display the parallel lines superimposed in the scatter plot.
3. Predict the excess body weight loss (EWL) of a 47 years old male if your PUID is an odd number otherwise predict the excess body weight loss (EWL) of 47 years old female.
4. Construct a 95% confidence interval and a prediction interval of the predicated value obtained in part (c).

*Solution:*

1. *We used R code below to import the data in R and the fitted multiple linear regression model is*

*EWL = 6.0755 + 0.4021 (age) for female*

*EWL= 2.6907 + 0.4021(age) for male*

*> Q2=read.csv("C:\\Users\\aryalg\\Desktop\\Q2.csv")*

*> head(Q2,3)*

*age gender EWL*

*1 49 0 24.2*

*2 54 1 25.4*

*3 37 0 20.1*

*> attach(Q2)*

*> model=lm(EWL~age+gender)*

*> model*

*Call:*

*lm(formula = EWL ~ age + gender)*

*Coefficients:*

*(Intercept) age gender*

*6.0755 0.4021 -3.3848*

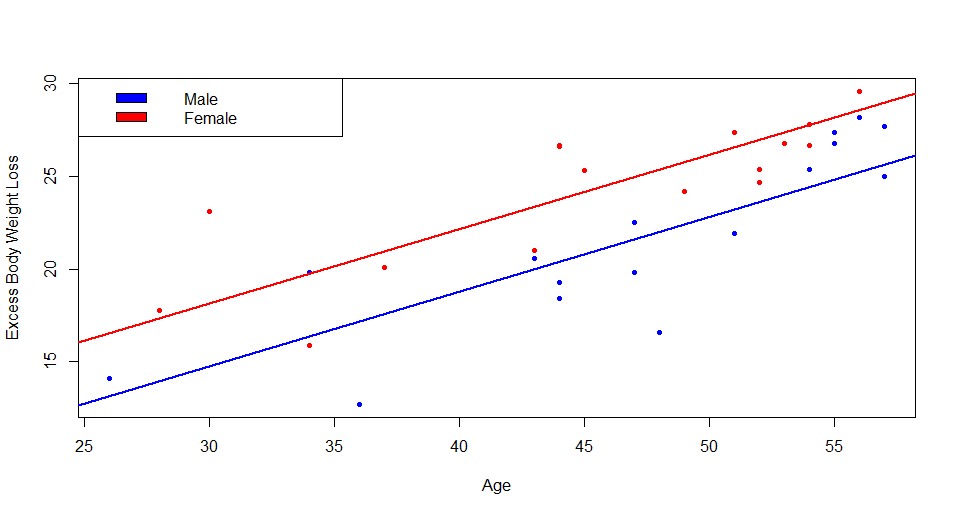
1. We used R code below to display the models

*> plot (age, EWL, xlab='Age',ylab='Excess Body Weight Loss', col = ifelse(gender == 1, 'blue', 'red'), pch = 20)*

*> legend("topleft", legend=c("Male", "Female"), fill=c("blue","red"))*

*> abline(6.0755,0.4021, lwd=2, col="red")*

*> abline(2.6907, 0.4021, lwd=2, col="blue")*



1. *We will make the prediction for female (gender=0) as the PUID is even. Based on the R output below the predicted excess body weight loss (EWL) for 47 years old female is 24.97*

*> predict(model,data.frame(age=47,gender=0))*

*1*

*24.97209*

*d) We used R code below to construct the confidence intervals.*

*> predict(model,data.frame(age=47,gender=0),interval='conf')*

*fit lwr upr*

*1 24.97209 23.70297 26.2412*

*The 95% confidence interval is (23.703, 26.241)*

*> predict(model,data.frame(age=47,gender=0),interval='pred')*

*fit lwr upr*

*1 24.97209 19.78149 30.16269*

*The 95% prediction interval is (19.78, 30.16)*

**Q. N. 3)** LTC stands for Long Term Care. Operators of long-term care homes are very interested in modeling their elderly residents are going to survive, because they need to plan. In one study, the variables for a sample of residents were: One year survival (1=Yes, 0=No), Age in years

Gender (1=F, 0=M), Indicator for dementia (1=Yes, 0=No)

The data is provided below

<https://www.utstat.toronto.edu/brunner/data/legal/ltc.data.txt>

a) Import the data in R and print the variable names.

b) Fit a simple logistic regression to model the alive as a response variable based on the age. Be sure to write the logistic regression model and display the fitted model in the scatterplot.

c) Update the above model by adding both gender and dementia variable.

Draw a random number between 0 and 9 using sample function.

d) Predict the likelihood of survival for a 82 years old female patient who is suffering from dementia if you have drawn an odd number and Predict the likelihood of survival for a 82 years old female patient who is not suffering from dementia if you have drawn an even number.

*Solution: a) We used R code below to import the data and print the variable names.*

*> Q3=read.table("https://www.utstat.toronto.edu/brunner/data/legal/ltc.data.txt", header=T)*

*> head(Q3)*

*age gender dementia alive*

*1 83 1 1 0*

*2 75 1 1 1*

*3 100 1 1 0*

*4 86 1 0 0*

*5 73 1 0 1*

*6 79 1 0 1*

*> names(Q3)*

*[1] "age" "gender" "dementia" "alive"*

b) Based on the R output below the fitted model is

*> fit=glm(alive~age, family=binomial)*

*> fit*

*Call: glm(formula = alive ~ age, family = binomial)*

*Coefficients:*

*(Intercept) age*

*22.2763 -0.2613*

*Degrees of Freedom: 246 Total (i.e. Null); 245 Residual*

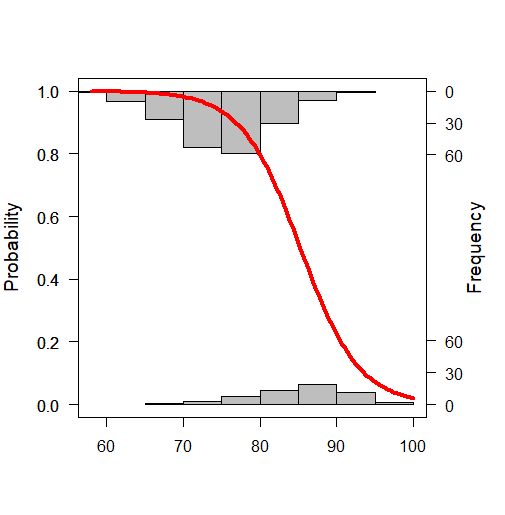
*Null Deviance: 266.9*

*Residual Deviance: 175.4 AIC: 179.4*

We can display the model using R code below

*> library(popbio)*

*>logi.hist.plot(age,alive,type="hist", boxp=F, col="gray" )*



*> newmodel=glm(alive~age+gender+ dementia, family=binomial)*

*> newmodel*

*Call: glm(formula = alive ~ age + gender + dementia, family = binomial)*

*Coefficients:*

*(Intercept) age gender dementia*

*23.3152 -0.2774 0.9672 -0.8745*

*Degrees of Freedom: 246 Total (i.e. Null); 243 Residual*

*Null Deviance: 266.9*

*Residual Deviance: 164.9 AIC: 172.9*

*> sample(0:9,1)*

*[1] 2*

1. *Based on the R output below, the likelihood of survival for a 82 years old female patient who is not suffering from dementia is 0.823*

*> predict(newmodel, data.frame(age=82, gender=1, dementia=0), type="resp")*

*1*

*0.8231218*

**Q.N. 4)** Office workers at a large insurance company are randomly assigned to one of 3 computers use training programs A, B and C) and their number of calls to IT support during the following month is recorded. Additional information on each worker includes years of experience and score on a computer literacy test (out of 100). The dataset is provided in the link below

<https://www.utstat.toronto.edu/brunner/data/legal/training.data.txt>

1. Import the data in R and determine its dimension.

If your last name has 5 or less letters, please answer (b) and (c) otherwise answer (d) and (e)

1. Display the computer literacy scores by training program using a parallel boxplot.
2. Test the hypothesis whether the score on computer literacy differ by the training program.
3. Display the Experience scores by training program using a parallel boxplot.
4. Test the hypothesis whether the experience scores differ by the training program

*Solution: We used R code below to import the data and it appears that there are 300 observations with 4 variables.*

*> Q4=read.table("https://www.utstat.toronto.edu/brunner/data/legal/training.data.txt", header=T)*

*> attach(Q4)*

*> dim(Q4)*

*[1] 300 4*

*> nchar("Aryal")*

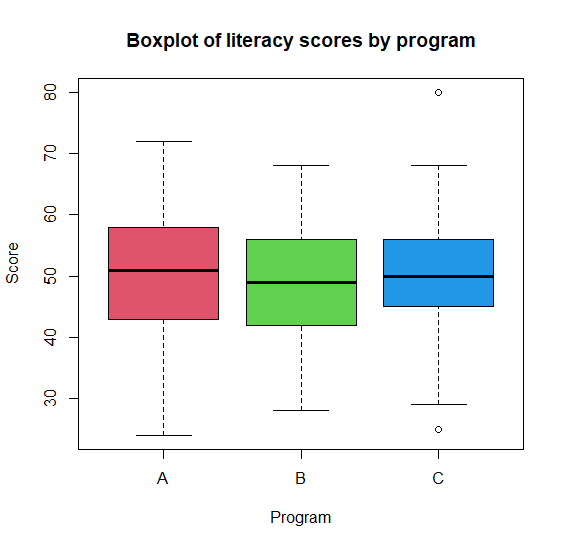
*[1] 5*

My last name has 5 letters so will answer (b) and (c)

*(b) We used R code below to display the scores by program*

*> boxplot(Score~Program)*

*> boxplot(Score~Program, col=c(2,3,4), main="Boxplot of literacy scores by program")*

**

*(c) Let denote that average literacy scores by program A, B and C respectively. We*

*would like to test the following hypothesis:*

*Based on the R output below, since p-value=0.564, we fail to reject the null hypothesis and have no evidence that the literacy score differs by program.*

*> summary(aov(Score~Program))*

*Df Sum Sq Mean Sq F value Pr(>F)*

*Program 2 109 54.61 0.573 0.564*

*Residuals 297 28286 95.24*

**Extra Credit:** Before the beginning of the Fall term, students in a first-year Calculus class took a diagnostic test with two parts: Pre-calculus and Calculus. Their High School Calculus marks and their marks in University Calculus were also available. In addition, other variables including gender, nationality are also provided in the link below. Not all students have complete data. Clean the data and determine the dimension of complete dataset.

<https://www.utstat.toronto.edu/brunner/data/legal/math.data.txt>

*Solution: Based on the output below there are 640 observations with 13 variables having complete information.*

*> EC=read.table("https://www.utstat.toronto.edu/brunner/data/legal/math.data.txt",na.strings=".")*

*> dim(EC)*

*[1] 1158 13*

*> Clean=na.omit(EC)*

*> dim(Clean)*

*[1] 640 13*