HW4_IS457_85

Mon Oct 8, 2018

In this assignment we will practice linear regression in R.

Part 1. Linear Regression Concepts (6pts)

These questions do not require coding, but you will need to explain the details. In this homework, "Regression" refers to the simple linear regression equation: y = b0 + b1*x

(1) (2pts) What is the interpretation of the coefficient b1? (What meaning does it represent?)

Ans :

In linear regression equation x is called the **explanatory** or **predictor** variable and y is called the **response** variable, b0 is referred to as **intercept** and b1 is referred as **coefficient**. b1 represents that for every **unit** increase in x value there will be a corresponding increase by b1 value in y.

(2) (2pts) Outliers are problems for many statistical methods but are particularly problematic for linear regression. Why is that? It may help to define what outlier means in this case. (Hint: Think of how residuals are calculated)

Ans :

Outliers are problems in linear regression because they can skew the regression line towards their direction thereby misclassifying other points which might be show some correlation among them. Typically, outliers that are horizontally far from the regression line have high leverage as they pull the regression line towards them harder thereby increasing the residual values for all the points. A residual value is the difference between the observed value and the value predicted by our model. Good models generally have low residual values. So, when a line skews towards the outliers they tend to increase the residual values of the points that might be more important to the problem domain.

(3) (2pts) How could you deal with outliers in order to improve the accuracy of your model?

Ans :

Outliers don’t need to be removed from the model to improve their performance. Models need to take into account exceptional cases in certain cases. For example, the models for market prediction need to take into the largest market swings – the “outliers” – in order to be valuable to financial firms. They need to be removed only after considering proper reasons. If the dataset includes a lot of outliers we could make use of a different method other than least squares to calculate the residuals as it gets affected greatly by the presence of outliers in the dataset.

Part 2. Sampling, Point Estimation, and creating functions

The following problems will use the Rabbit dataset and explore the Blood Pressure change(BPchange) for Rabbit in control group of "Treatment". Load the data by running the following code:

library(MASS)

data(Rabbit)

(4) (1) Subset the data frame to include ONLY rabbits (observations) in control group of "Treatment". (2pts) Name it 'rabbitCon' and show the first 10 observations of your output.(2pts)

Ans :

rabbitCon = Rabbit[Rabbit$Treatment=="Control",]

head(rabbitCon,n = 10)

```
BPchange   Dose Run Treatment Animal
1      0.50   6.25  C1   Control     R1
2      4.50  12.50  C1   Control     R1
3     10.00  25.00  C1   Control     R1
4     26.00  50.00  C1   Control     R1
5     37.00 100.00  C1   Control     R1
6     32.00 200.00  C1   Control     R1
7      1.00   6.25  C2   Control     R2
8      1.25  12.50  C2   Control     R2
9      4.00  25.00  C2   Control     R2
10    12.00  50.00  C2   Control     R2
```

The output above shows that the first ten rows of the **rabbitCon** dataset.

(4) (2) Use the sample function to generate a vector of 1s and 2s with the same length as rabbitCon, call it 'group'.(2pts) Use this vector to split the 'BPchange' variable into two vectors, BP_V1 and BP_V2. (4pts) Print out the vectors group, BP_V1, BP_V2 and the lengths of BP_V1 and BP_V2.

IMPORTANT: Make sure to run the seed function before running the sample function to ensure the result is reproducible.

set.seed(457) # DO NOT change

Ans :

group = sample(c(1,2),nrow(rabbitCon),replace = TRUE)

temp = split(rabbitCon$BPchange, group)

BP_V1 = temp[[1]]

BP_V2 = temp[[2]]

BP_V1

```
[1]  4.50 10.00 26.00 32.00  1.25  4.00 12.00  0.75  3.00  3.00
[11] 14.00 24.00 33.00  1.50  1.50  5.00 16.00 18.00
```

BP_V2

```
[1]  0.50 37.00  1.00 27.00 29.00 22.00  1.25  1.50  6.00 19.00
[11] 33.00 20.00
```

length(BP_V1)

```
[1] 18
```

length(BP_V2)

```
[1] 12
```

(5) (1) Calculate the mean and the standard deviation for each of the two vectors, BP_V1 and BP_V2. (4pts) Create a 95% confidence interval for your sample means using Z score.(4pts) (you can use the following formula for the Confidence Interval: mean +/- 2 * standard deviation). Compare the confidence intervals, do they seem to agree or disagree, explain (their ranges? differences?). (2pts)

Ans :

mean(BP_V1)

```
[1] 11.63889
```

sd(BP_V1)

```
[1] 10.93437
```

mean(BP_V2)

```
[1] 16.4375
```

sd(BP_V2)

```
[1] 13.72379
```

BP_V1_Low_CI = mean(BP_V1) - 1.96 * (sd(BP_V1)/length(BP_V1)^0.5)

BP_V1_Low_CI

```
[1] 6.587467
```

BP_V1_Up_CI = mean(BP_V1) + 1.96 * (sd(BP_V1)/length(BP_V1)^0.5)

BP_V1_Up_CI

```
[1] 16.69031
```

BP_V2_Low_CI = mean(BP_V2) - 1.96 * (sd(BP_V2)/length(BP_V2)^0.5)

BP_V2_Low_CI

```
[1] 8.672537
```

BP_V2_Up_CI = mean(BP_V2) + 1.96 * (sd(BP_V2)/length(BP_V2)^0.5)

BP_V2_Up_CI

```
[1] 24.20246
```

From the above output we can see that the two confidence variable seem to agree as the mean values for both the group tend to fall into the confidence interval. The confidence interval for the BP_V2 group seems to be wider than the BP_V1 interval which is reasonable as the differences in the mean values of both the groups is more as well.

(5) (2) From what you practice in 5 (1), let's generalize the calculation process. (5pts) Write a function to calculate the 95% confidence intervals of any input vector (numerical) x, according to the formula given in the previous question.

Ans :

```
CI_95 = function(v)
    {
     m = mean(v)
     s = sd(v)
     n = length(v)
     lci = m - 1.96*s/n^0.5
     hci = m + 1.96*s/n^0.5
     return(c(lci,hci))
    }
```

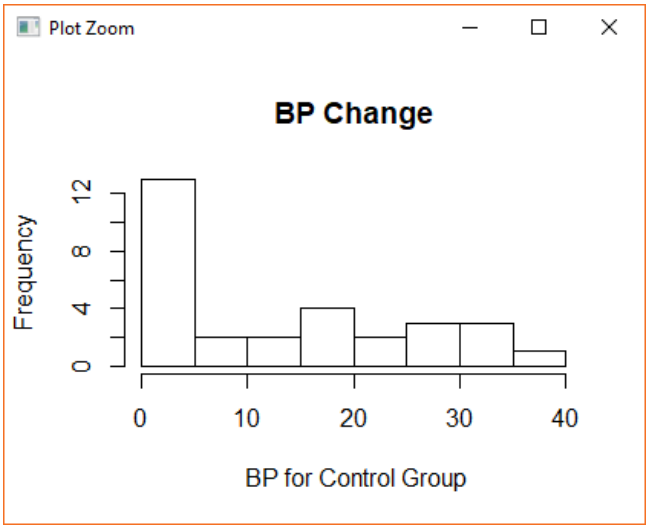
```
CI_95(BP_V1)
```

```
[1]  6.587467 16.690310
```

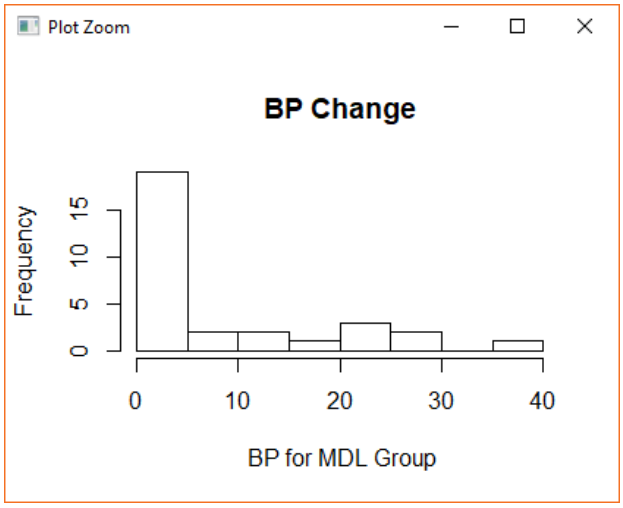
(6) Using the hist() function, plot a histogram of BPchange of rabbits under control group as well as for the MDL group (separately). (2pts) Do the histograms resemble a normal distribution? why or why not? (2pts) Comment on the shape of the distributions you see in the histograms. What does the shape indicate in the context of this dataset?(4pts)

Ans :

```
hist(Rabbit$BPchange[Rabbit$Treatment=="Control"])
```



hist(Rabbit$BPchange[Rabbit$Treatment=="MDL"])



Both the histograms do not resemble normal distribution but rather a rightly skewed distribution. In the context of this dataset it can be said most of the rabbits from both the groups have a low blood pressure and also the value of dose values in both the groups are rightly skewed and tend to have some affect on BP values of rabbits. These values of the BP are certainly not normally distributed and are rightly skewed for both the groups.

Part 3. Linear Regression

This problem will use the same dataset as Part 2. We will focus on two variables:
BP change: change in blood pressure relative to the start of the experiment.
Dose: dose of Phenylbiguanide in micrograms.
To start with, let us define a null hypthosesis. If we want to test the effect of dosage on BPchange, the null hypthosesis is:

H0: Dosage has no effect on BPchange.
H0: B1 = 0
HA: B1 != 0

(7) Fit a linear regresssion using Dose to predict BPchange, using lm() for rabbits under MDL treatment. (2pts) Name it 'model_BP'. What function would you use to get the summary statistics from lm models? Go ahead and use it. (2pts) Examine the model diagnostics using plot(). Comment on the plots, what do the fitted values, noise, outliers look like?(8pts) Would you consider this a good model or not? Please explain. (2pts)

Ans :

rabbitMDL = Rabbit[Rabbit$Treatment=="MDL",]

model_BP = lm(BPchange~Dose, data = rabbitMDL)

summary(model_BP)

plot(model_BP)

The residuals vs fitted plot shows us the residual values for each of the points and their distance from the regression line. We can see a lot of outliers in this plot. The regression line does pass through some of the observed values and we do see some cloud of points near the regression line for low dose values. Also, the variability of the observed values increases from the regression line as the value of dose increases. There are influential outliers both above and below the regression line. Based on these observations a linear model is not suitable for this variable as although the variable clearly has a strong positive linear trend as the they have a correlation value of 0.90 and the observations are independent but due to lack of constant variability across the regression line and not normal residuals makes it difficult to model all the observed values.

(8) With the summary statistics from above, calculate the 95% confidence interval for Dose using t score (2pts) Note: use this code to find the t score: tvalue <- qt(1-0.05/2,nrow(rabbitMDL)-2)

Ans :

```
tvalue <- qt(1-0.05/2,nrow(rabbitMDL)-2)

m_MDL = mean(rabbitMDL$Dose)

s_MDL = sd(rabbitMDL$Dose)

lci_MDL = m_MDL - tvalue * (s_MDL/(nrow(rabbitMDL))^0.5)

hci_MDL = m_MDL + tvalue * (s_MDL/(nrow(rabbitMDL))^0.5)

ci_MDL = c(lci_MDL,hci_MDL)
```

(9) Based on the result from Q7& Q8 (p-value and CI), would you reject the null hypothesis or not? Explain. (2pts)

Ans :

Based on the results in Q7 & Q8 we would reject the null hypothesis as the p-value 1.159e-11 is very less than the 0.05 level of significance value. So, the dosage has some effect on the Blood Pressure of the rabbits.