STAT 40001/STAT 59800 **Statistical Computing** **Fall 2020**

Homework 4-Solution

Due : November 10 , 2020

Name:
PUID:

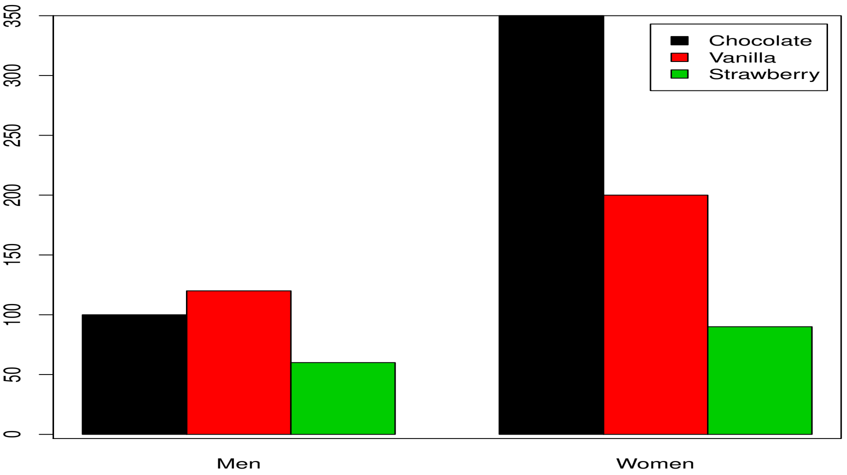
Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw output containing ERRORs). Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations.

Q.N. 1) An ice-cream store is interested to determine whether or not there is an association between gender and preference for ice cream flavor. Data below provides the information based on the order received in a randomly chosen day. Perform an appropriate analysis to check whether the ice cream flavor and gender are independent.

	Flavor		
	Chocolate	Vanilla	Strawberry
Men	100	120	60
Women	350	200	90

Solution: First we will display the data using side-side barplot

```
> Men=c(100,120, 60)
> Women=c(350,200,90)
> names(Men)=c("Chocolate", "Vanilla", "Strawberry")
> names(Women)=c("Chocolate", "Vanilla", "Strawberry")
> icecream=cbind(Men,Women)
> barplot(icecream, beside=T, col=c(1,2,3), legend=rownames(icecream))
> box()
```



We would like to test whether the gender and ice cream flavor preferences are independent using Chi-squared test.
$\boldsymbol{H_0}$: *Gender and ice-cream flavors are independent*
$\boldsymbol{H_1}$: *Gender and ice-cream flavors are not independent*

```
> icecream
           Men Women
Chocolate  100   350
```

```
Vanilla    120   200
Strawberry  60    90
> chisq.test(icecream)

        Pearson's Chi-squared test

data:  icecream
X-squared = 28.362, df = 2, p-value = 6.938e-07
```

Decision: Since p-value is small, we reject the null hypothesis and conclude that Gender and ice-cream flavors are not independent.

Q.N. 2) A clinical dietician wants to compare two different diets, A and B, for diabetic patients. She hypothesizes that diet A (Group 1) will be better than diet B (Group 2), in terms of lower blood glucose. She plans to get a random sample of diabetic patients and randomly assign them to one of the two diets. At the end of the experiment, which lasts 6 weeks, a fasting blood glucose test will be conducted on each patient. She also expects that the average difference in blood glucose measure between the two group will be about 10 mg/dl. Furthermore, she also assumes the standard deviation of blood glucose distribution for diet A to be 15 and the standard deviation for diet B to be 17. How many subjects are needed in each group assuming equal sized groups? (Please use $\boldsymbol{\alpha} = \mathbf{0.05}$ and Power=0.8).

Solution: Since what really matters is the difference, instead of means for each group, we can enter a mean of zero for Group 1 and 10 for the mean of Group 2, so that the difference in means will be 10. Next, we need to specify the pooled standard deviation, which is the square root of the average of the two standard deviations squared (i.e., variances). In this case, it is $\sqrt{\frac{\mathbf{15^2+17^2}}{\mathbf{2}}} = \mathbf{16.03}$. *Therefore,*

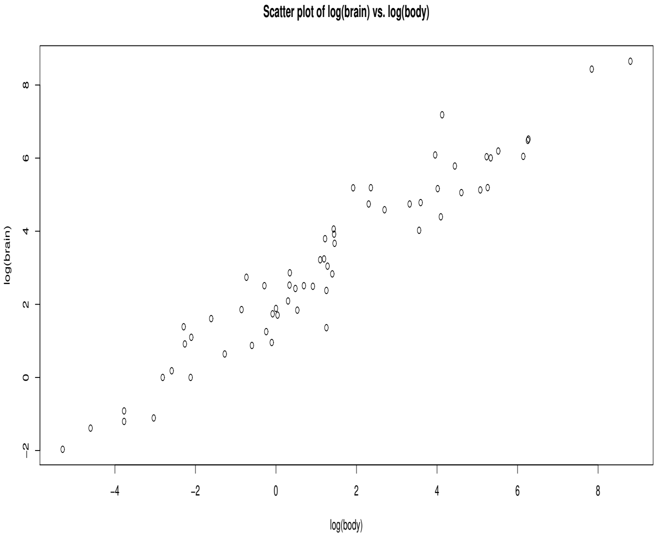
```
> library(pwr)
> pwr.t.test(d=(0-10)/16.03,power=.8,sig.level=.05,type="two.sample", alt="two.sided")

     Two-sample t test power calculation

              n = 41.31968
              d = 0.6238303
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
NOTE: n is number in *each* group
```

Based on the R output, it is required that there are at lest 42 individuals in each group.

Q.N. 3) The mammals data set in the MASS package records brain size and body size of 62 different mammals.
a) Display a scatter plot of the log(brain) vs. log(body).
b) Fit a simple linear regression model to the transformed data.
c) What is the equation of the fitted model.

Solution:
a) Please note that we will be using natural ***log*** *i.e* ***ln*** *. It should be noted that log in R means ln*
We will use the R code below to convert the data and plot the scatter plot of the data

```
> library(MASS)
> attach(mammals)
> x1=log(body)
> y1= log(brain)
> plot(x1,y1, xlab="log(body)", ylab="log(brain)", main="Scatter plot of log(brain) vs. log(body)")
```



b) Fit a simple linear regression model to the transformed data.
Solution: We can use R code below to fit the regression model

```
> library(MASS)
> attach(mammals)
> x1=log(body)
> y1= log(brain)
> > model=lm(y1~x1)
> model
lm(formula = y1 ~ x1)
Coefficients:
(Intercept)           x1
     2.1348       0.7517
```

Hence, the fitted model is

$$\boldsymbol{log(brain) = 2.1348 + 0.7517 \times log(body).}$$

c) What is the equation of the fitted model.
Solution: We have the fitted model $\boldsymbol{log(brain) = 2.1348 + 0.7517 \times log(body)}$. *which can be simplified as below*

$$\begin{aligned}
\boldsymbol{log(brain)} &= \boldsymbol{2.1348 + 0.7517 \times log(body)} \\
\boldsymbol{log(brain) - 0.7517 \times log(body)} &= \boldsymbol{2.1348} \\
\boldsymbol{log\left(\frac{brain}{(body)^{0.7517}}\right)} &= \boldsymbol{2.1348} \\
\boldsymbol{\frac{brain}{(body)^{0.7517}}} &= \boldsymbol{e^{2.1348}} \\
\boldsymbol{brain} &= \boldsymbol{8.455 \times (body)^{0.7517}}
\end{aligned}$$

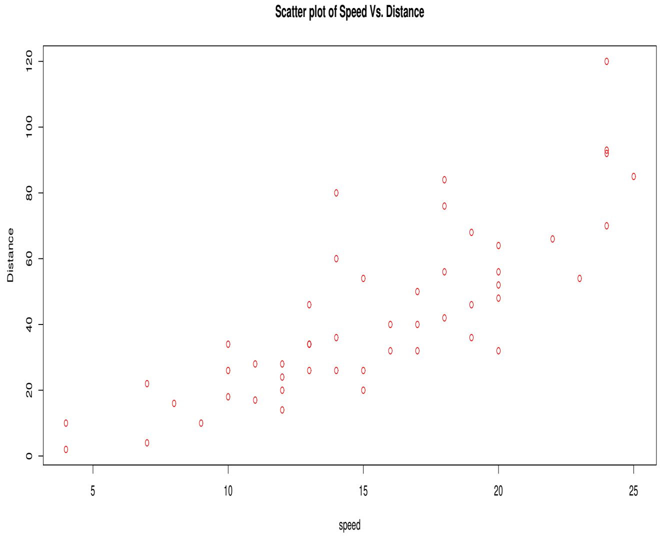
Therefore, the equation of the fitted model in the original unit of measurements is

$$\boldsymbol{brain = 8.455 \times (body)^{0.7517}}$$

Q.N. 4) The data set cars is one of the data sets installed with R and is available in base package. The data set contains 50 observations of speed(mph) and dist(stopping distance in feet).
a) Display the data using scatter plot.
b) Fit a simple regression model using speed as a predictor variable.
c) Add the fitted line to the scatter plot.
d) Calculate the residuals and fitted values and print only first five observations of the residuals and fitted values.
e) Create a scatter plot of the residuals and fitted values.
f) Assuming that no intercept model is appropriate fit a simple linear regression model.
g) Calculate and compare the coefficient of determination for both the with intercept and no-intercept models.
h) Using your fitted model predict the stopping distance for a car with an speed of 21 mph.

a) Display the data using scatter plot.
Solution: We will read and plot the scatter plot of the data using R code below:

```
> data=cars
> attach(cars)
> names(cars)
[1] "speed" "dist"
> plot(speed,dist, main= "Scatter plot of Speed Vs. Distance", ylab="Distance", col=2)
```



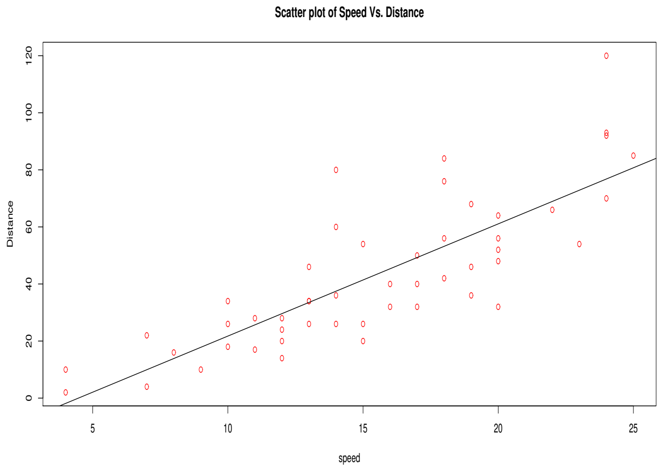
b) Fit a simple regression model using speed as a predictor variable.
Solution: Use R code below to estimate the parameters

```
> model=lm(dist~speed)
> model
Call:
lm(formula = dist ~ speed)
Coefficients:
(Intercept)        speed
    -17.579        3.932
```

Therefore, the fitted model is $\boldsymbol{distance = -17.579 + 3.932 \times speed}$

c) Add the fitted line to the scatter plot.
Solution: We can add the fitted line to the scatter plot using the code below:

```
> plot(speed,dist, main= "Scatter plot of Speed Vs. Distance", ylab="Distance", col=2)
> model=lm(dist˜speed)
> abline(model)
```

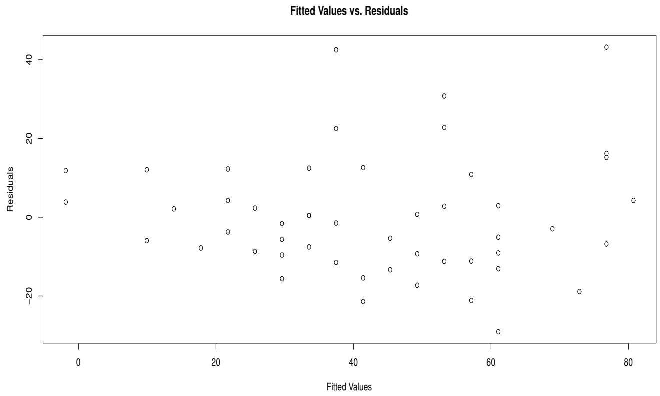


d) Calculate the residuals and fitted values and print only first five observations of the residuals and fitted values.
Solution: The fitted value and residuals of the model are calculated using the R code below:

```
> model=lm(dist˜speed)
> fitted=fitted(model)
> residuals=resid(model)
> head(fitted,5)
        1         2         3         4         5 
-1.849460 -1.849460  9.947766  9.947766 13.880175 
> head(residuals,5)
        1         2         3         4         5 
 3.849460 11.849460 -5.947766 12.052234  2.119825 
```

e) Create a scatter plot of the residuals and fitted values.
Solution: R code below is used to create the scatter plot of the fitted value and residuals.

```
> model=lm(dist˜speed)
> fitted=fitted(model)
> residuals=resid(model)
> plot(fitted,residuals,xlab="Fitted Values",ylab="Residuals",main="Fitted Values vs. Residuals")
```



f) Assuming that no intercept model is appropriate fit a simple linear regression model.
Solution: No-intercept model can be fitted using the R code below:

```
> model1=lm(dist~-1+speed)
> model1
Call:
lm(formula = dist ~ -1 + speed)
Coefficients:
speed
2.909
```

Hence, the fitted model is $\boldsymbol{Distance = 2.909 \times Speed}$
g) Calculate and compare the coefficient of determination for both the with intercept and no-intercept models.
Solution: In order to calculate the coefficient of determination we use the R code below:

```
> summary(model)
> summary(model1)
```

Note that we have the following values for the coefficient of determination:

Model	R^2	AdjustedR^2
Intercept Model	0.6511	0.6438
No-intercept	0.8963	0.8942

h) Using your fitted model predict the stopping distance for a car with an speed of 21 mph.

Solution: We can use R code below to predict the stopping distance for a car with an speed of 21 mph

```
> model=lm(dist~speed)
> model1=lm(dist~-1+speed)
> xval=as.data.frame(21)
> colnames(xval)="speed"
> predict(model,xval)
       1
65.00149
> predict(model1,xval)
       1
61.09178
```

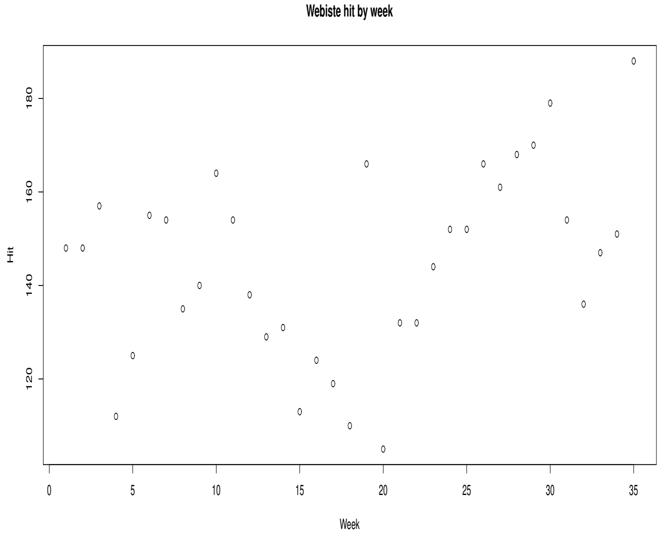
Note that the intercept model predicts a stopping distance of 65.00149 feet and the no-intercept model predicts a stopping distance of 61.09178 feet.

Q.N. 5) An author maintains a website on a particular book and using Google Analytics, records the number of visits on this particular website on each day of the year. As expected there are more hits during weekdays then on weekends. Since the book is used as a textbook for a statistics course there are more hits during the time when the classes are in session. Table below provides the data for 35 weeks from April through November 2009. To explore the week by week visit patterns of these

Week	Hits
1	148
2	148
3	157
4	112
5	125
6	155
7	154
8	135
9	140
10	164
11	154
12	138
13	129
14	131
15	113
16	124
17	119
18	110
19	166
20	105
21	132
22	132
23	144
24	152
25	152
26	166
27	161
28	168
29	170
30	179
31	154
32	136
33	147
34	151
35	188

a) Display the data using a scatterplot.
b) Calculate the rank correlation coefficient to measure the association between the week and the number of hits on the website.
c) Test for the significance of the correlation at **0.05** level.
Solution:
a) We imported the data in the R readable format and plotted the scantier plot using R code below

```
> data=read.table("C://STAT 40001//Data sets//website.txt", header=T)
> plot(data, main="Scatter Plot of the Website Hits")
```



b) Calculate the rank correlation coefficient to measure the association between the week and the number of hits on the website.
Solution: Use R code below to calculate the rank correlation coefficient

```
> data=read.table("C://STAT 40001//Data sets//website.txt", header=T)
> cor.test(data$Week,data$Hits, method="spearman")

        Spearman's rank correlation rho

data:  data$Week and data$Hits
S = 4842.713, p-value = 0.05945
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3217489
```

Therefore, the rank correlation coefficient is 0.3217489.
c) Test for the significance of the correlation at **0.05** level.
Solution: We would like to test the hypothesis

$$\boldsymbol{H_0} \;:\; \boldsymbol{\rho = 0}$$
$$\boldsymbol{H_a} \;:\; \boldsymbol{\rho \neq 0}$$

From the R output in part (b) we see that p-value = 0.05945 which is greater than 0.05, so we fail to reject the null hypothesis. This means we don't have enough evidence to say that the week and amount of website hits are correlated.