BIST/STAT 5605 Homework 1

Due date: 11:59pm, Thursday, February 6, 2026

General Instructions

- Use this R Markdown template for homework submission.
- Answer the questions by inserting R code and necessary comments if applicable. Your output must contain the R code (do not use the `echo=FALSE` option) if applicable.
- Save the compiled PDF file under the file name `LastName-FirstName-HW1.pdf` and submit it through HuskyCT by the deadline.

Possible Max Points: 45 points

Each problem, should it be selected for grading, will be worth 5 points except for problems 6 and 7; problems 6 and 7 will each be worth 15 points. Each problem, which is not selected for grading, will be worth 2 points for completion.

(1) (5 or 2 points) When asked to state the simple linear regression model, a student wrote as follows: $E[Y_i] = \beta_0 + \beta_1 X_i + \epsilon_i$. Do you agree?

(2) (5 or 2 points) In a simulation exercise, regression model on page 19 of note 1 applies with $\beta_0 = 100$, $\beta_1 = 20$, and $\sigma^2 = 25$. An observation on Y will be made for $X = 5$.

(a) Can you state the exact probability that Y will fall between 195 and 205? Explain.

(b) If the normal error is assumed, can you now state the exact probability that Y will fall between 195 and 205? If so, state it.

(3) (5 or 2 points) The regression function relating production output by an employee after taking a training program (Y) to the production output before the training program (X) is $E\{Y\} = 20 + 0.95X$, where X ranges from 40 to 100. An observer concludes that the training program does not raise production output on the average because β_1 is not greater than 1.0. Comment.

(4) (5 or 2 points) Evaluate the following statement: "For the least squares method to be fully valid, it is required that the distribution of Y be normal."

(5) (5 or 2 points) According to page 36 of note 1, $\sum_{i=1}^n e_i = 0$ when a SLR model is fitted to a set of n cases by the method of least squares. Is it also true that $\sum_{i=1}^n \epsilon_i = 0$? Comment.

(6) (15 or 2 points) The least squares regression line for a given set of data with a sample size of $n = 20$ is $\hat{Y} = -42 + 0.9X$ (i.e., $b_0 = -42$ and $b_1 = 0.9$). The MSE of the fitted simple linear regression (SLR) model is 0.14, and the standard error of b_1 (i.e., $se(b_1)$) is 0.016. Suppose $\bar{X} = 200$. Answer the following questions and additionally provide references for the pertinent equation numbers from the notes and/or textbook.

(a) What is the fitted value of Y at $X = 220$.

(b) Compute the standard error of b_0.

(c) Find $\bar{Y}$.

(d) What are S_{XX} and S_{XY} for this data set?

(e) Compute $\text{corr}(b_0, b_1)$.

(7) (15 or 2 points) Suppose you are given n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

(a) Describe an empirical Q-Q plot and a scatter plot for this data set?

(b) Can an empirical Q-Q plot be identical to the respective scatter plot for certain data set? If so, when would this happen?

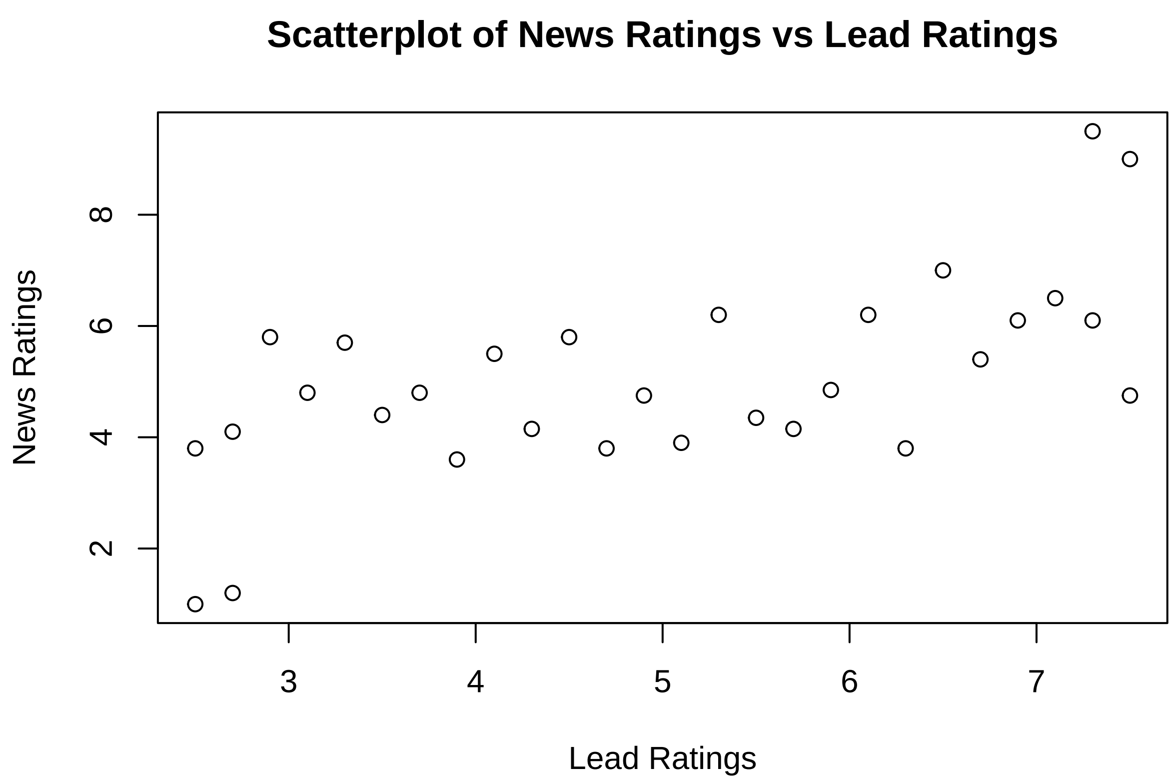
(c) To gain better understanding between these two types of plots, draw the empirical Q-Q plot and the scatter plot for the Ratings of TV Shows Data in Example 2 from the HuskyCT class website. Provide a brief discussion. Data file:

```
setwd("M:/teaching/stat5605s26/homework")
ratings = read.csv("ratings.csv")
head(ratings)
```

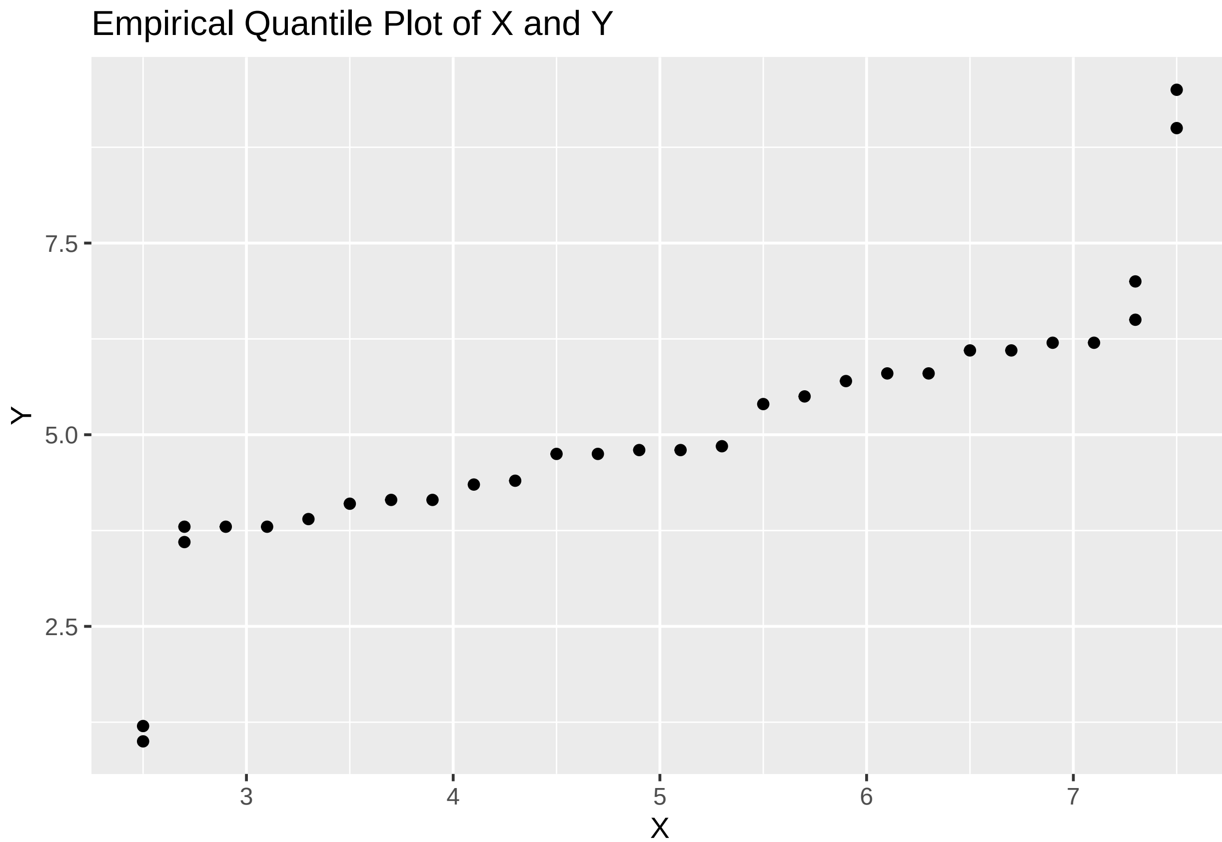
```
##     X   Y
## 1 2.5 3.8
## 2 2.7 4.1
## 3 2.9 5.8
## 4 3.1 4.8
```

```
## 5 3.3 5.7
## 6 3.5 4.4
```

```
attach(ratings)
library(ggplot2)
plot(X,Y, main="Scatterplot of News Ratings vs Lead Ratings",
     ylab="News Ratings", xlab="Lead Ratings")
```




```
sx <- sort(X); sy <- sort(Y)
lenx <- length(sx); leny <- length(sy)
ggplot() + geom_point(aes(x = sx, y = sy)) +
ggtitle("Empirical Quantile Plot of X and Y") +
xlab("X") + ylab("Y")
```



(8) (5 or 2 points) Suppose you are given n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let e_i denote the residual for the i^{th} observation calculated based on the Least Squares method. Using algebra of least squares, argue that weighted sum of residuals, with i^{th} residual weighted by the corresponding Y_i, is SSE.

(9) (5 or 2 points) A student was investigating from a large sample whether variables Y_1 and Y_2 follow a bivariate normal distribution. The student obtained the residuals when regressing Y_1 on Y_2, and also obtained the residuals when regressing Y_2 on Y_1, and then prepared a normal probability plot for each set of residuals. Do these two normal probability plots provide sufficient information for determining whether the two variables follow a bivariate normal distribution? Explain.

(10) (5 or 2 points) The data below show, for a consumer finance company operating in seven cities, the number of competing loan companies operating in a city (X_i) and the number per thousand of delinquent loans made in that city (Y_i):

X_i	4	1	2	3	3	4	2
Y_i	18	4	9	14	16	20	8

For a simple linear regression analysis, let X denote the design matrix and Y denote the column vector of responses for the dataset in reference above. Compute $X'X$, $X'Y$, $(X'X)^{-1}$, and use these results to find the estimated vector $\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$ of the regression coefficients.

```
one=rep(1,7)
x1=c(4, 1, 2, 3, 3, 4, 2)
X=t(rbind(one,x1))
X
```

```
##      one x1
## [1,]   1  4
## [2,]   1  1
## [3,]   1  2
## [4,]   1  3
## [5,]   1  3
## [6,]   1  4
## [7,]   1  2
```

```
Y=c(18, 4, 9, 14, 16, 20, 8)
Y
```

```
## [1] 18  4  9 14 16 20  8
```

```
XTY=t(X) %*% Y
XTY
```

```
##     [,1]
## one   89
## x1   280
```