STAT 5605 Homework 1

February 5, 2026

Jack Bienvenue

Contents

1 Problem 1

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?

I do **not** agree with the student's statement of the simple linear regression model. While it is very close to being correct, there is a conceptual mistake. In the correct simple linear regression (SLR) model, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$, we have a few components:

1. The value of the outcome variable for a specific observation, Y_i,
2. The SLR model intercept, β_0,
3. The SLR model slope, β_1,
4. The value of the input variable for a specific observation, X_i,
5. and the random error associated with the observation's outcome variable value, ϵ_i.

The student incorrectly added an expectation function around Y_i. This is incorrect as the *expectation* of random variable Y_i does not include the random error associated with the realization. The expectation of Y_i given X_i **under our SLR model** is actually $E[Y_i|X_i] = \beta_0 + \beta_1 X_i$.

2 Problem 2

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 $X = 5$.

2.1 (a)

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

Let's begin by building our SLR model using the coefficients given by the problem description:

$$Y = 100 + (20 \times 5) + \epsilon_i, \text{with assumptions:}$$

1. Expectation of errors is 0, i.e. $E(\epsilon_i) = 0$,
2. Homoscedasticity, i.e. $Var(\epsilon_i) = \sigma^2$,
3. Errors are uncorrelated between observations.

In this case, we *cannot* state the exact probability that Y will fall between 195 and 205 because *although we have all relevant information for the important parameters* $(\beta_0, \beta_1, Var(\epsilon_i))$ *in our model*, we do not have information about the specific distributional shape of the errors (ϵ_i), disallowing us to make statements about the exact probability of an observation's Y value to fall in a given interval.

2.2 (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.

```
# Calculate the probabilities
## P(Y<=195)
prob_195 <- pnorm(195, mean = 200, sd = 5)
## P(Y<=205)
prob_205 <- pnorm(205, mean = 200, sd = 5)
## P(Y<=205) - P(Y<=195) = P(195 <= Y <= 205)
prob_bw_195_205 <- prob_205 - prob_195
cat("Probability of Y being in [195, 205]:", prob_bw_195_205)
```

```
## Probability of Y being in [195, 205]: 0.6826895
```

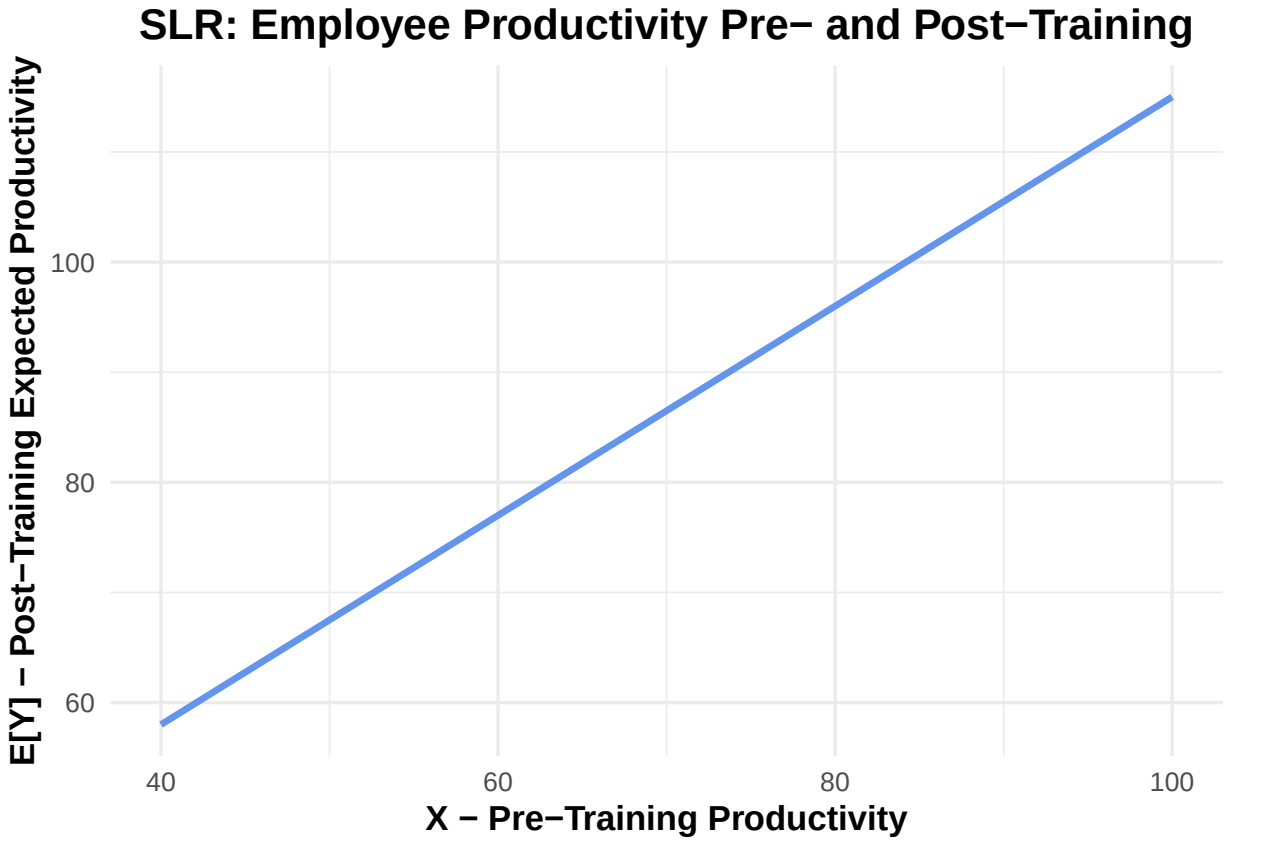
If we assume that errors are normally distributed, $\epsilon_i \sim N(0, \sigma^2)$, we are now able to calculate probabilities for Y falling in certain intervals. These intervals are "exact" under a normal error ($N(0, \sigma^2)$) assumption, and will be exact if this assumption accurately describes the error distribution (otherwise, the calculated probability will be an approximation).
To calculate the probability that Y falls between 195 and 205 given that $\beta_0 = 100$, $\beta_1 = 20$, and $X = 5$, we should first use this information to obtain $E[Y_{X=5}]$, the "center" for our errors: $Y = 100 + (20 \times 5) = 200$. From here, we can use R's `pnorm()` function, setting the mean to 200 (our mean-zero errors are centered at $E[Y_{X=5}]$) and standard deviation to 5 (our *variance* for the errors is 25, therefore the relevant standard deviation is $\sqrt{25} = 5$). We can calculate the lower-tail-to-quantile probabilities and subtract the lower probability (corresponding to 195) from the higher probability (corresponding to 205) to find the probability of Y falling in the 195-205 range. Based on the empirical rule, we expect this probability to be around 68% because the bounds are one standard deviation removed from the mean. We do find this result to be true in our calculation, as the computed probability is 0.6826895. This probability is exact if the normal assumption on the errors we made is true and if all parameter values were the true population parameters. Otherwise, this probability is approximate.

3 Problem 3

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.

```
# Create plot
x <- seq(40, 100, by = 1)
y <- 20 + 0.95 * x
```

```r
df <- data.frame(x = x, y = y)
ggplot(df, aes(x = x, y = y)) +
  geom_line(linewidth = 1.2, color = "cornflowerblue") +
  labs( title = "SLR: Employee Productivity Pre- and Post-Training",
    x = "X - Pre-Training Productivity",
    y = "E[Y] - Post-Training Expected Productivity" ) +
  theme_minimal(base_size = 13) +
  theme( plot.title = element_text(face = "bold", hjust = 0.5),
    axis.title = element_text(face = "bold"))
```



While the observer may be wary of the efficacy of the training program due to the fact that $\beta_1 < 1$, we can assure them that there is a positive effect of the training program on employee productivity on average in this linear model. When we plot our regression line, we can easily observe that employees across all pre-training productivity levels experienced an increase in productivity after undergoing training, on average using this model. Even in such a simple linear model, the regression line is being defined by two different parameters, β_0 and β_1, and therefore the effect is being "split" between these contributors. The *combination* of $\beta_0 = 20$ and $\beta_1 = 0.95$ actually yields a regression that would suggest that, on average, there is a positive productivity effect associated with undergoing training. A coefficient $\beta_1 < 1 \nRightarrow E[Y|X] < X$, because algebraically, $20 + 0.95X > X$ for $X \in [40, 100]$.

4 Problem 4

Evaluate the following statement: "For the least squares method to be fully valid, it is required that the distribution of Y be normal."

5 Problem 5

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 Problem 6

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.

6.1 (a)

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

6.2 (b)

Compute the standard error of b_0.

6.3 (c)

Find $\bar{Y}$.

6.4 (d)

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

6.5 (e)

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

7 Problem 7

Suppose you are given n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

7.1 (a)

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

7.2 (b)

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

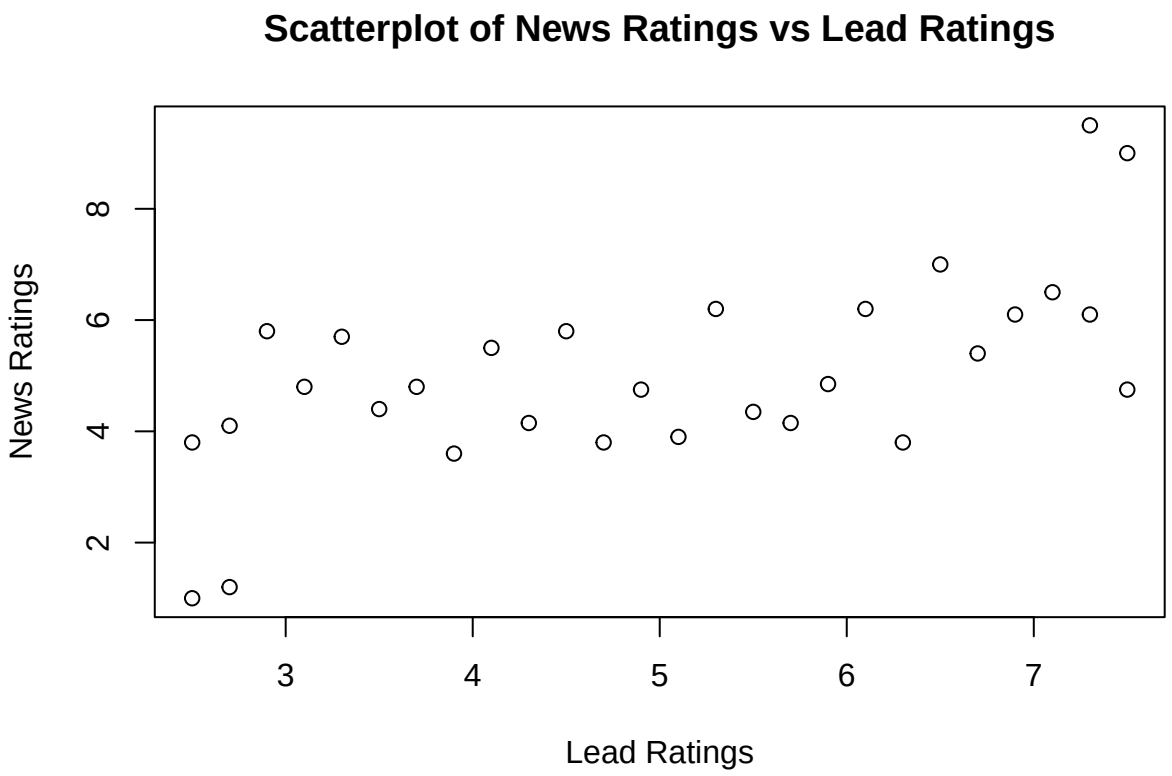
7.3 (c)

To gain better understanding betweenthese 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.

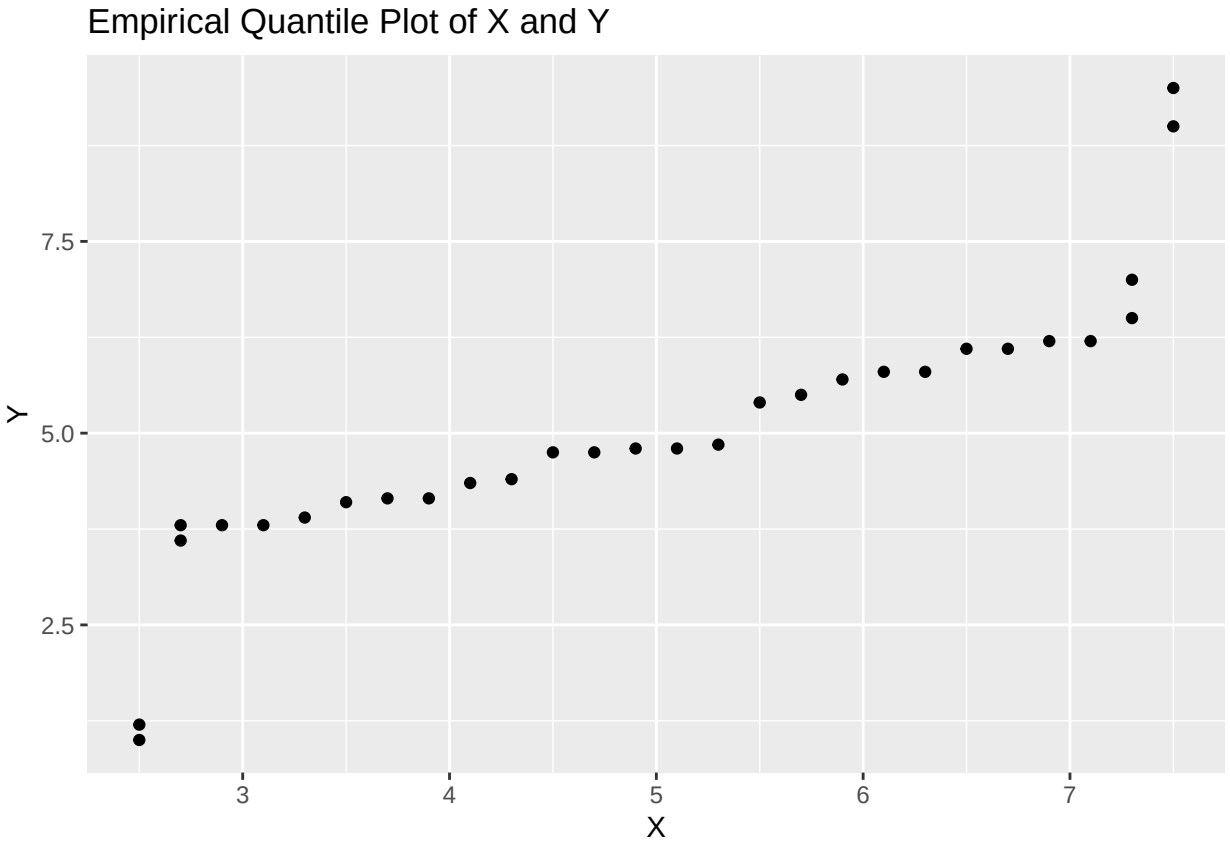
```
ratings = read.csv("ratings.csv")
head(ratings)
```

```
##     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 Problem 8

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 Problem 9

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 Problem 10

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
```