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
```

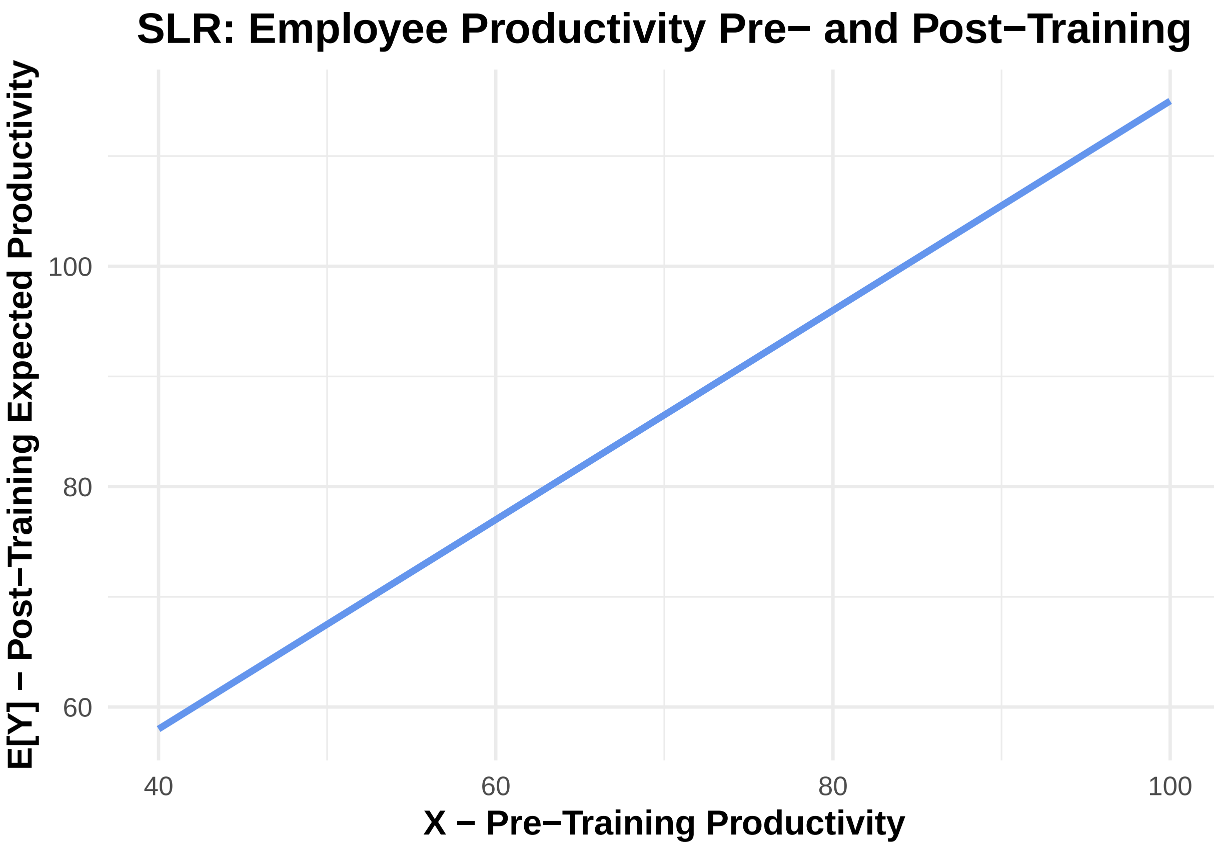
If we assume that errors are normally distributed, $\epsilon_i \sim N(0, \sigma^2)$, we are now able to calculate probabilities for Y falling in certain intervals. These intervals are "exact" under a normal error ($N(0, \sigma^2)$) assumption, and will be exact if this assumption accurately describes 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 $\beta_1 < 1 \not\Rightarrow E[Y|X] < X$, 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."

This statement is **false** in general. The method of constructing a least squares (LS) fit is built on linear algebra and calculus that does not require any distributional assumptions for any of the structural components (X, $\mathbf{Y}$, $\mathbf{e}_i$, etc.). So, normality in the true distribution of the response variable is not required for generating valid model coefficient estimates (which are calculated under $\mathrm{E}(\epsilon_i) = 0$ and homoscedasticity assumptions). If one wanted to perform inference with the resulting coefficients attained from the LS fit, normality would become important for the validity of that operation. The LS method in general does not universally require an assumption of normality of the response Y.

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.

No, it is not true. The fact that $\sum_{i=1}^{n} e_i = 0$ after LS fitting for an SLR model is a consequence of the method itself, as it is necessary for minimizing the sum of squared residuals during the estimation process. The true errors, ϵ_i, are not able to be observed and are random variables (in contrast, residuals are not random, and are determined entirely by the fitted model and observed data). The true random errors for the true model will average to a value increasingly close to 0 under many distributions of ϵ as n, the number of observations, approaches infinity, justifying the assumption for SLR, but the true errors need not sum to zero always. In fact, it would be incredibly surprising if the true errors summed to exactly 0 under most circumstances.

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 $X = 220$.

The fitted value of Y at $X = 220$ can be easily calculated using the given fitted model:

$$\hat{Y}_{X=220} = -42 + 0.9 \times 220 = 156$$

6.2 (b)

Compute the standard error of b_0.

We can compute the standard error of $\hat{\beta}_0$ using some facts we know about the composition of the MSE and the SE for $\hat{\beta}_1$ alongside the formula for $SE[\hat{\beta}_0]$ itself:

$$\text{MSE} = \frac{1}{20-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = s^2 = 0.14 \quad \text{with } s \text{ the residual standard deviation,} \tag{1}$$

$$\text{SE}\left(\hat{\beta}_1\right) = 0.016 = \frac{s}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}, \tag{2}$$

$$\text{SE}\left(\hat{\beta}_0\right) = s\sqrt{\frac{1}{20} + \frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}} \tag{3}$$

After substituting known values, we attain:

$$s = \sqrt{MSE} = \sqrt{0.14} = 0.37416;$$

$$\text{SE}\left(\hat{\beta}_1\right) = 0.016 = \frac{0.37416}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}} \therefore \sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \equiv \sqrt{S_{xx}} = \frac{0.37416}{0.016} = 23.385,$$

$$\text{As a result, } S_{xx} = (23.385)^2 = 546.875, \text{and returning to } \hat{\beta}_0:$$

$$\text{SE}[\hat{\beta}_0] = (0.37416)\sqrt{\frac{1}{20} + \frac{\bar{X}^2}{S_{xx}}} = (0.37416)\sqrt{\frac{1}{20} + \frac{(200)^2}{546.875}} = 3.201$$

FOOTNOTE:
Equation (1) comes from textbook (Applied Linear Statistical Models (hereafter ALSM) Fifth Edition, Kutner et al.) Equation 1.22; Equation (2) comes from adapting ALSM Equation 2.3b, Equation (3) is adapted from ASLM Equation 2.22b.

6.3 (c)

Find $\bar{Y}$.

The mean response of observations, $\bar{Y}_{sample}$, can be identified from the fitted model and our known value of $\bar{X}$. In an SLR model, the point $\left(\bar{X}, \bar{Y}\right)$ appears on the line*, and each value of X has unique corresponding $\hat{Y}$. We can solve our model for $\bar{Y}$ using $\bar{X}$:

$$\bar{Y}_{sample} = -42 + 0.9\bar{X} = -42 + 0.9(200) = 138$$

**FOOTNOTE:* Fact comes from Slide 35 of Note 1.

6.4 (d)

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

The quantity S_{XX} is equivalent to $\sum_{i=1}^{n} X_i^2 - n\bar{X}^2$; the quantity S_{XY} is equivalent to $\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$ according to Note 1, page 33.
In language, S_{XX} is a measure of the variability of X around its sample mean and S_{XY} is a measure of the magnitude of X and Y's variance together around their respective sample means.
We previously calculated S_{XX} in Question 6.b using the known values given in the question and the equation for the standard error of the slope coefficient estimate. We found $S_{XX} = 546.875$ there.
Now, S_{XY} can be calculated by exploiting the relationship between S_{XY} and two known values: $\hat{\beta}_1$ and S_{XX}. The SLR LS solution slope $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$. So,

$$S_{XY} = \hat{\beta}_1 S_{XX} = 0.9(546.875) = 492.1875$$

6.5 (e)

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

The correlation of the fitted parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ can be calculated using the formula given on page 45 of Note 1:

$$\text{corr}(b_0, b_1) = \frac{-\sigma^2 \frac{\bar{X}}{S_{XX}}}{\left[\sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right) \sigma^2 \frac{1}{S_{XX}}\right]^{1/2}} = \frac{-\bar{X}}{\left[\frac{S_{XX}}{n} + \bar{X}^2\right]^{1/2}},$$

Which we can easily use by substituting known values for $\bar{X}$, S_{XX} and n:

$$\text{corr}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-200}{\left[\frac{546.875}{20} + 200^2\right]^{1/2}} = -0.99965,$$

Allowing us to see a very strong negative correlation present between the sample intercept and slope for this SLR model.

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?

EMPIRICAL Q-Q Plot:

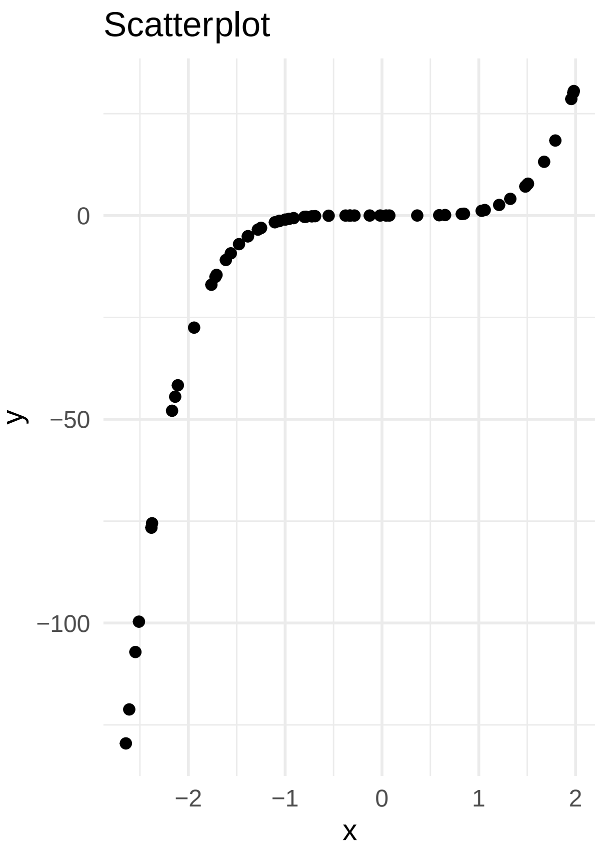
An empirical Q-Q plot, or empirical quantile-quantile plot, is a two-dimensional graph which plots pairs of ordered points. Since the natural pairing of the observations is discarded, we instead get a sense of how the distribution of two variables differs (and we implicitly control for different centering of the distributions). Variables with distributions with very similar shapes (but not necessarily centers) will combine in a Q-Q plot to form a chain of points that appears like a straight 45 degree line. Different behavior of the shape of the chain of points plotted in a Q-Q plot, like curvature (overall and towards ends), parallel displacement from the ideal line, rotation from the line, can all help to diagnose differences in distribution for issues like differences in tail behavior, distributional location, spread, etc.

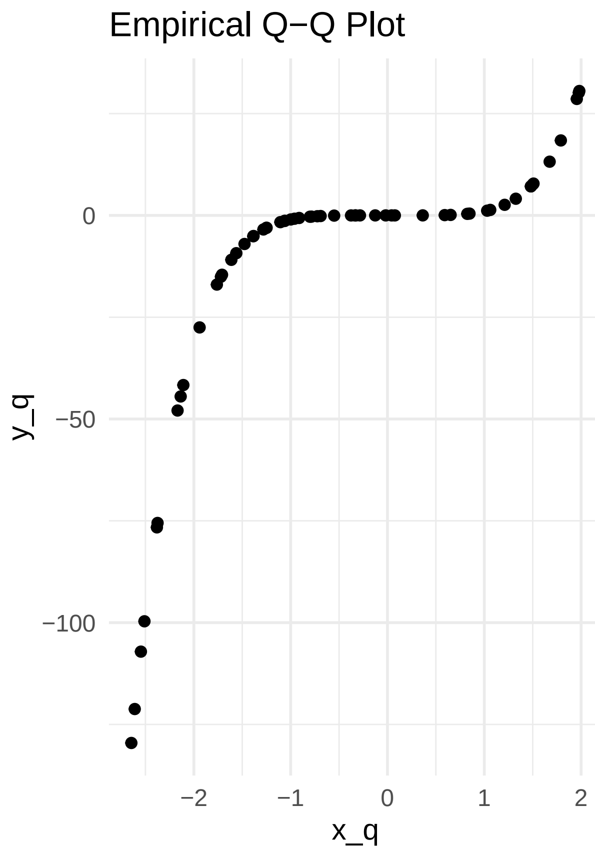
SCATTERPLOT:
Simpler than the Q-Q plot, the scatterplot simply displays the X- and Y-coordinates of paired observations of the variables. This plot primarily serves the purpose of giving the viewer a sense of correlation (or other relationship) between two variables, checking for linearity/non-linearity, and dispersion of points along an imagined line of best fit.

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?

```r
# Example of coinciding Q-Q- and scatter-plot
library(ggplot2)
library(gridExtra)
set.seed(5605)
x <- sort(runif(58, -3, 2))
y <- x^5
df <- data.frame(x = x, y = y)
p_scatter <- ggplot(df, aes(x = x, y = y)) +
  geom_point() +
  ggtitle("Scatterplot") +
  theme_minimal()
df_qq <- data.frame(
  x_q = sort(df$x),
  y_q = sort(df$y))
p_qq <- ggplot(df_qq, aes(x = x_q, y = y_q)) +
  geom_point() +
  ggtitle("Empirical Q-Q Plot") +
  theme_minimal()
grid.arrange(p_scatter, p_qq, ncol = 2)
```





COINCIDENCE:

In general, the empirical Q-Q plot and scatterplot do not coincide. However, it is possible for the plots to coincide. Specifically, this will occur when the observations used for generating the Q-Q plot (individually re-ordered pairs) and original pairs from observations coincide. This can happen if observations are relatively increasing in both X and Y simultaneously without either variable "jumping" ahead in a way such that it would be reordered under Q-Q plot constructing. A sufficient situation to cause the Q-Q plot to be identical to the scatter plot would be when $|\text{corr}(X, Y)| = 1$, but all that is required is *perfect rank correlation* between X and Y, meaning that the orders of X and Y observations are the same or *exactly* opposite.
The example in the plot displays a scenario when the ordering of X and Y do not change under ranking, leading to identical Q-Q- and scatter-plots.

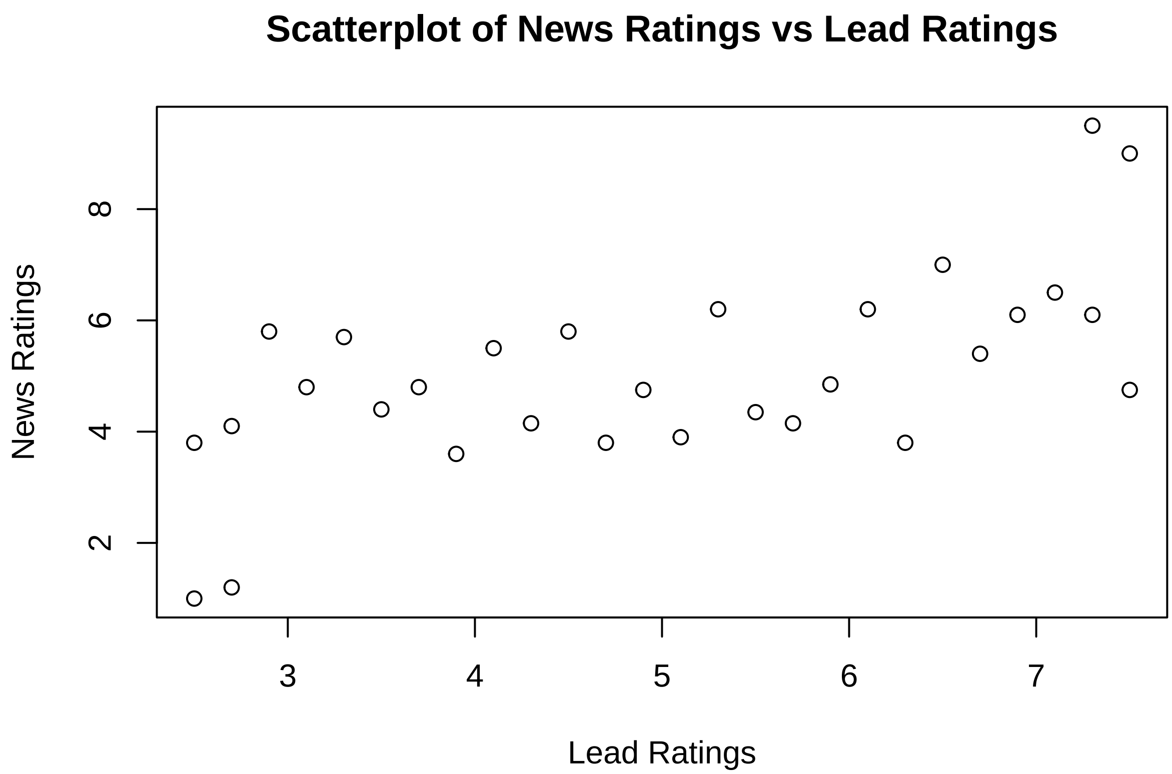
7.3 (c)

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.

```
ratings = read.csv("data/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
```

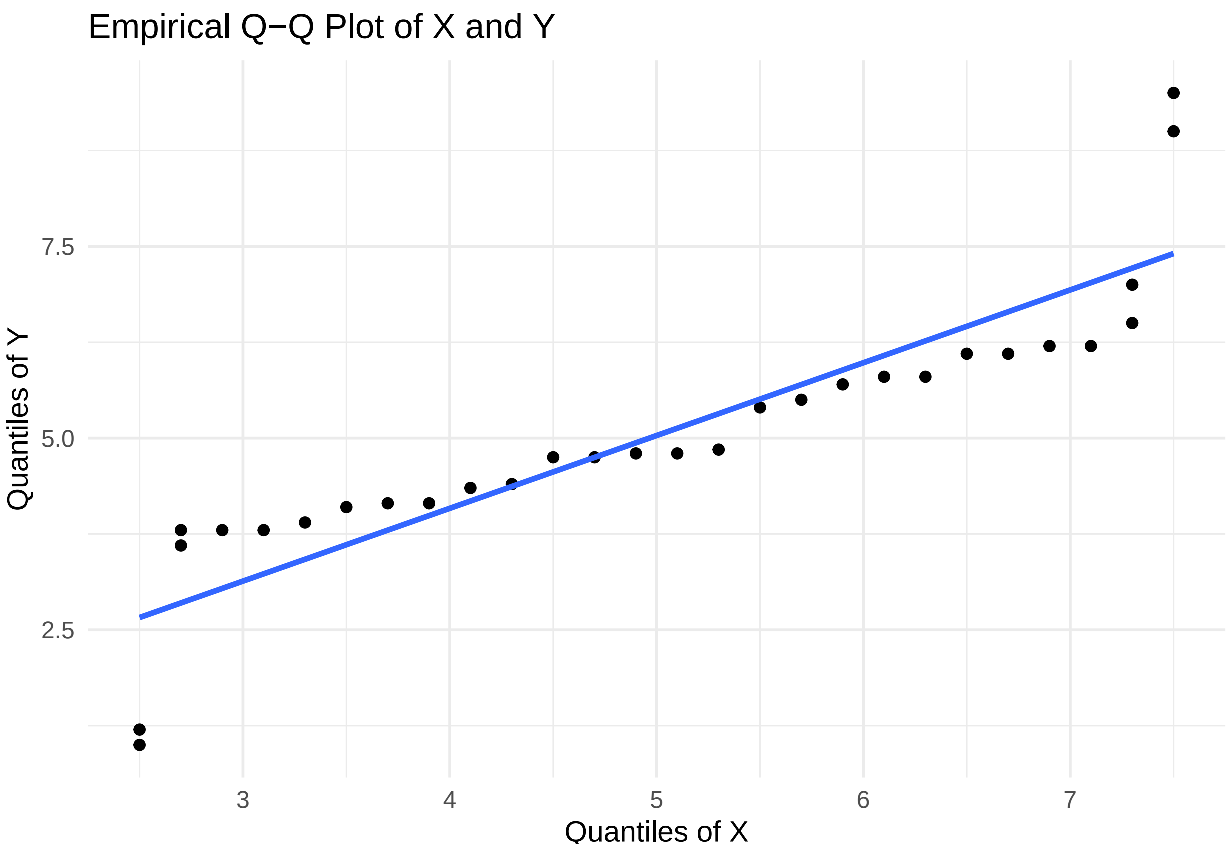
```
attach(ratings)
library(ggplot2)
plot(X,Y, main="Scatterplot of News Ratings vs Lead Ratings",
     ylab="News Ratings", xlab="Lead Ratings")
```




```
sx <- sort(ratings$X)
sy <- sort(ratings$Y)
qq_df <- data.frame(sx = sx, sy = sy)
ggplot(qq_df, aes(x = sx, y = sy)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1) +
  labs(
    title = "Empirical Q-Q Plot of X and Y",
    x = "Quantiles of X",
    y = "Quantiles of Y"
```

```
) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Unlike the example I provided for the previous sub-question, we do not observe identical plots. This time, in our scatterplot, we observe a relationship that appears mostly linear (with some potentially nonlinear behavior in the tails), with a potentially slightly positive slope. We observe that we do not have any sort of perfectly deterministic relationship, and we observe a moderate degree of dispersion of the points relative to the scale of the variables.
In the Q-Q Plot, we observe some departure from the ideal, matched-quantile line. As X's quantiles grow, the path of points crosses the line to hover below it, indicating a shift from the quantiles of Y falling above those of X to the opposite. We observe some unusual tail behavior as well, with the quantiles of Y located far below those of X for the smallest values and far above those of X at the largest values.
In terms of the show ratings, we observe a possible positive correlation between lead show rating and a trend that could possibly be modeled with an SLR model in our scatterplot. We are not given any strong graphical evidence to dispute the presence of a holdover effect. In our Q-Q plot, we observe some non-linearity and divergent tail behavior, implying that the distribution of the lead-in show's ratings and the subject show's ratings are not identical.

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.

The claim given above is that:

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} Y_i e_i,$$

In our linear models, we know that the value of any Y_i can be decomposed into a fitted value and residual, i.e. $Y_i = \hat{Y}_i + e_i$.
We can make a substitution into the prior equation using this relationship:

$$\text{SSE} = \sum_{i=1}^{n} (\hat{Y}_i + e_i) e_i = \left(\sum_{i=1}^{n} \hat{Y}_i e_i \right) + \left(\sum_{i=1}^{n} e_i^2 \right),$$

Recalling that fitted values and errors are orthogonal, the dot product being calculated in the $\left(\sum_{i=1}^{n} \hat{Y}_i e_i \right)$ term reduces to 0. Therefore, we verify that $\sum_{i=1}^{n} Y_i e_i$ indeed reduces to 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.

The setup that the student has is insufficient for determining whether the two variables follow a bivariate normal distribution. The student will have information about the marginal distributions of Y_1 and Y_2 and about the conditional distributions of $Y_1|Y_2$ and $Y_2|Y_1$. However, even if these distributions are all normal (which is necessary), that does not necessarily imply joint normality of (Y_1, Y_2).

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.

```r
# DESIGN MATRIX
cat("DESIGN MATRIX X (with intercept):\n")
```

```
## DESIGN MATRIX X (with intercept):
```

```r
one <- rep(1, 7)
x1  <- c(4, 1, 2, 3, 3, 4, 2)
X <- t(rbind(one, x1))
print(X)
```

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

```r
# RESPONSE VECTOR
cat("\nRESPONSE VECTOR Y:\n")
```

```
##
## RESPONSE VECTOR Y:
```

```r
Y <- c(18, 4, 9, 14, 16, 20, 8)
print(Y)
```

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

```r
# MATRIX CALCULATIONS

## X'X
cat("\nX'X:\n")
```

```
##
## X'X:
```

```r
XtX <- t(X) %*% X
print(XtX)
```

```
##     one x1
## one   7 19
## x1   19 59
```

```r
## X'Y
cat("\nX'Y:\n")
```

```
##
## X'Y:
```

```r
XtY <- t(X) %*% Y
print(XtY)
```

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

```r
## (X'X)^(-1)
cat("\n(X'X)^(-1):\n")
```

```
##
## (X'X)^(-1):
```

```r
XtX_inv <- solve(XtX)
print(XtX_inv)
```

```
##            one         x1
## one  1.1346154 -0.3653846
```

```
## x1  -0.3653846  0.1346154
```

```r
## Estimated beta vector
cat("\nEstimated regression coefficients (b0, b1):\n")
```

```
##
## Estimated regression coefficients (b0, b1):
```

```r
b <- XtX_inv %*% XtY
print(b)
```

```
##          [,1]
## one -1.326923
## x1   5.173077
```

The implementation of the matrix algebra is complete above in the code section. The attained estimated vector is:

$$\vec{\hat{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} -1.326923 \\ 5.173077 \end{pmatrix},$$

This was calculated using $\vec{\hat{\beta}} = (X'X)^{-1}X'Y$ from Note 1, page 42.