Problem 1. The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.

a. Describe the relationship between volume and height of these trees.

b. Describe the relationship between volume and diameter of these trees.

c. Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

a. moderate, positive, linear relationship
   - non-constant variance (as x ↑, var ↑)

b. strong, positive, linear/slightly curved
   - constant variance

c. diameter - stronger relationship
Problem 2. The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).

a. Describe the relationship between distance and travel time.
b. How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers?
c. The correlation between travel time (in miles) and distance (in minutes) is \( r = 0.636 \). What is the correlation between travel time (in kilometers) and distance (in hours)?

\begin{itemize}
\item[a.] weak, positive \( \rightarrow \) potentially linear
\item[b.] It wouldn’t change \( \rightarrow \) our units don’t affect form, direction, or strength of a relationship.
\item[c.] \( r = 0.636 \) \( \text{limits don’t affect correlation} \)
Problem 3. What would be the correlation between the annual salaries of males and females at a company if for a certain type of position men always made:

a. $5,000 more than women?

b. twice as much as women?

c. 25% less than women?

\[ \text{a. } \text{pay}_M = \text{pay}_W + 5000 \]
\[ r = 1 \]

\[ \text{b. } \text{pay}_M = 2(\text{pay}_W) \]
\[ r = 1 \]

\[ \text{c. } \text{pay}_M = 0.75(\text{pay}_W) \]
\[ r = 1 \]
Problem 4. Determine if the following statements are true or false. Explain

a. A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation of 0.5.

b. Correlation is a measure of the association between any two variables.

a. True - strength is determined by magnitude of $r$. -0.90 is further from 0 than 0.50.

b. False - correlation is a measure of the linear association between any two numerical variables.
Problem 5. Match each correlation to the corresponding scatterplot.

a. \( r = -0.72 \) - Scatterplot 2
b. \( r = 0.07 \) - Scatterplot 4

C. \( r = 0.86 \) - Scatterplot 1

d. \( r = 0.99 \) - Scatterplot 3
Problem 6. In college freshman men, it appears as though there is a linear relationship between height (in inches) and weight (in pounds). In a sample of the population, we see that the average height is 68.4 inches, with a standard deviation of 4.0 inches. We see that the average weight is 141.6 pounds, with a standard deviation of 9.6 pounds. The correlation between height and weight in our sample is 0.73. We would like to create a linear model that can be used to predict a male college freshman’s weight, given we know their height.

a. What is the formula for the linear model?

\[ \hat{y} = b_0 + b_1 x \]

b. Interpret the slope.

goto through: \((\bar{x}, \bar{y})\)

\[ y \text{- weight} + x \text{- height} \]

c. Interpret the intercept.

d. What is r-squared?

e. Interpret r-squared.

f. James is a male college freshman who is 68 inches tall. What is his predicted weight?

g. James actually weighs 152 pounds. What is his residual?

a. \[ b_1 = \left( \frac{S_y}{S_x} \right) r = \left( \frac{9.6}{4.0} \right) (0.73) = 1.752 \]

\[ b_0 : \quad \bar{y} = b_0 + b_1 \bar{x} \]

\[ 141.6 = b_0 + 1.752(68.4) \]

\[ 141.6 = b_0 + 119.8368 \]

\[ b_0 = 141.6 - 119.8368 \]

\[ b_0 = 21.7632 \]

\[ \hat{y} = 21.7632 + 1.752x \]
Problem 6. In college freshman men, it appears as though there is a linear relationship between height (in inches) and weight (in pounds). In a sample of the population, we see that the average height is 68.4 inches, with a standard deviation of 4.0 inches. We see that the average weight is 141.6 pounds, with a standard deviation of 9.6 pounds. The correlation between height and weight in our sample is 0.73. We would like to create a linear model that can be used to predict a male college freshman’s weight, given we know their height.

a. What is the formula for the linear model?
b. Interpret the slope.
c. Interpret the intercept.
d. What is r-squared?
e. Interpret r-squared.
f. James is a male college freshman who is 68 inches tall. What is his predicted weight?
g. James actually weighs 152 pounds. What is his residual?

b. Slope (b1): For each additional inch in height, our predicted weight increases by 1.752 pounds.

c. Slope (b0): For someone who is 0 inches tall, we predict their weight to be 21.7632 pounds. (Pretty meaningless)
Problem 6. In college freshman men, it appears as though there is a linear relationship between height (in inches) and weight (in pounds). In a sample of the population, we see that the average height is 68.4 inches, with a standard deviation of 4.0 inches. We see that the average weight is 141.6 pounds, with a standard deviation of 9.6 pounds. The correlation between height and weight in our sample is 0.73. We would like to create a linear model that can be used to predict a male college freshman’s weight, given we know their height.

a. What is the formula for the linear model?
b. Interpret the slope.
c. Interpret the intercept.
d. What is r-squared?
e. Interpret r-squared.
f. James is a male college freshman who is 68 inches tall. What is his predicted weight?
g. James actually weighs 152 pounds. What is his residual?

d. \[ r^2 = (0.73)^2 = 0.5329 \]

e. 53.29\% of the variation in weight can be explained by our model (height).
Problem 6. In college freshman men, it appears as though there is a linear relationship between height (in inches) and weight (in pounds). In a sample of the population, we see that the average height is 68.4 inches, with a standard deviation of 4.0 inches. We see that the average weight is 141.6 pounds, with a standard deviation of 9.6 pounds. The correlation between height and weight in our sample is 0.73. We would like to create a linear model that can be used to predict a male college freshman’s weight, given we know their height.

a. What is the formula for the linear model?

b. Interpret the slope.

c. Interpret the intercept.

d. What is r-squared?

e. Interpret r-squared.

f. James is a male college freshman who is 68 inches tall. What is his predicted weight?

g. James actually weights 152 pounds. What is his residual?

\[
\hat{y} = 21.7632 + 1.752 \times x
\]

\[
\hat{y} = 21.7632 + (1.752)(68)
\]

\[
\hat{y} = 21.7632 + 119.136
\]

\[
\hat{y} = 140.8992 \text{ pounds}
\]

\[
\text{residual} = y_i - \hat{y}_i
\]

\[
= 152 - 140.8992
\]

\[
= 11.1008 \text{ pounds}
\]
Problem 7. The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch) and the percentage of bike riders in the neighborhood wearing helmets (helmet). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%.

a. If the r-squared for the least-squares regression line for the data is 72%, what is the correlation between the two variables?
b. What is the least squares regression line?
c. Interpret the intercept of the least-squares regression line.
d. Interpret the slope of the least-squares regression line.
e. What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual.

\[ r^2 = 0.72 \]
\[ r = \sqrt{0.72} \]
\[ r = \pm 0.849 \]

Based on scatterplot: \( r = -0.949 \)
Problem 7. The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch) and the percentage of bike riders in the neighborhood wearing helmets (helmet). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%.

\[
\hat{y} = b_0 + b_1 x
\]

a. If the r-squared for the least-squares regression line for the data is 72%, what is the correlation between the two variables?

b. What is the least squares regression line?

c. Interpret the intercept of the least-squares regression line.

d. Interpret the slope of the least-squares regression line.

e. What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual.

\[
b_0 = \bar{y} - b_1 \bar{x} = 38.8 - (-0.537)(30.8) = 55.34
\]

\[
\hat{y} = 55.34 - 0.537 x
\]
Problem 7. The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch) and the percentage of bike riders in the neighborhood wearing helmets (helmet). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%.

Intercept: 55.34
Slope: -0.537

a. If the r-squared for the least-squares regression line for the data is 72%, what is the correlation between the two variables?
b. What is the least squares regression line?
c. Interpret the intercept of the least-squares regression line.
d. Interpret the slope of the least-squares regression line.
e. What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual.

C. For a neighborhood where 0% receive reduce-fee lunch, we predict 55.34% of bike riders to wear a helmet.

d. For each additional percentage point of children who receive reduced-fee lunch, we predict a decrease of 0.537% in the percent of bike riders who wear a helmet.
Problem 7. The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch) and the percentage of bike riders in the neighborhood wearing helmets (helmet). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%.

a. If the r-squared for the least-squares regression line for the data is 72%, what is the correlation between the two variables?
b. What is the least squares regression line?
c. Interpret the intercept of the least-squares regression line.
d. Interpret the slope of the least-squares regression line.
e. What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual.

\[ y = 55.34 - 0.537x \]
\[ y = 55.34 - (0.537)(40) \]
\[ y = 33.86 \]

Residual: \( y - \hat{y} = 40 - 33.86 \approx 6.14 \)

- Positive value
- Actual is higher than predicted
- Prediction: underestimate
Problem 8. Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America’s two-party system, one political theory suggests the higher the unemployment rate, the worse the President’s party will do in the midterm elections. To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2018, with the exception of those elections during the Great Depression. We consider the percent change in the number of seats of the President’s party (e.g., percent change in the number of seats for Republicans in 2018) against the unemployment rate. Below, you are given both a scatterplot of the data as well as the regression output. You want to test and see if there is a linear relationship at the 0.10 level.

![Scatterplot with regression output]

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| Intercept| -7.3644    | 5.1553  | 1.43     | 0.1646   |
| unemp    | -0.8807    | 0.8350  | -1.07    | 0.2961   |

a. What kind of relationship do you notice?
b. The data for the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively. Do you agree that they should be removed for this investigation? Why or why not?
c. What is the least squares regression line?
d. What are the hypotheses?
e. What is the significance level?
f. What is the value of the test statistic?
g. What is the p-value?
h. What is the correct decision?
i. What is the appropriate conclusion/interpretation?
j. What is the 95% confidence interval for the slope parameter?
a. Weak-moderate relationship
   negative relationship

b. These are high leverage points
   potentially influential points
   - we are leaving out data

c. $\hat{y} = -7.3644 - 0.8897 x$

d. $H_0: \beta_1 = \Phi$
   $H_A: \beta_1 \neq \Phi$

e. $\alpha = 0.10$
\[ f_0 \quad TS = \frac{\text{point estimate - null value}}{\text{standard error}} \]

\[ TS = \frac{-0.8897 - 0}{0.8350} \]

\[ TS = \frac{-0.8897}{0.8350} = -1.066 \]

g. \quad TS = -1.066

\[ df = 27 \]

\[ 0.057 < |TS| < 1.314 \]

\[ 0.204 \text{ p-value} < 0.30 \]
\[ h. \quad 0.20 < p\text{-value} < 0.30 \]
\[ p\text{-value} > 0.20 > 0.10 \]
\[ p\text{-value} > 0.10 \]
\[ p\text{-value} > \alpha \]

\[ \boxed{\text{Fail to Rej. } H_0} \]

i. The data does not provide statistically significant evidence that there is a linear relationship between unemployment and midterm election results.
j. point est ± margin of error

\[ b_1 \pm t^* (\text{Std. Error}) \]

\[-0.8897 ± (2.052)(0.8350) \]

\[-0.8897 ± 1.71342 \]

\[-0.8897 - 1.71342 = -2.60 \]

\[-0.8897 + 1.71342 = 0.82 \]

95% CI: (-2.60, 0.82)
Problem 9. A researcher wanted to see if she could predict the number of sentences in an advertisement based on the number of words. She collected data on this using a sample of size 54. Below, you are given both a scatterplot of the data as well as the regression output.

![Scatterplot and Regression Output](image)

<table>
<thead>
<tr>
<th>Parameter estimates:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Slope</td>
</tr>
</tbody>
</table>

a. What kind of relationship do you notice between number of words and number of sentences?
b. What is the least squares regression line?
c. What is the estimate for the slope parameter? How would we interpret this?
d. What is the estimate for the intercept parameter? How would we interpret this?
e. The value for r-squared is 0.57. How do we interpret this value? What is the value of the correlation coefficient?
f. What is the 95% confidence interval for the slope parameter?
g. Interpret your interval from part f.

**Solution:**

- a. *Moderate, linear, positive relationship*
- b. $\hat{y} = 5.39 + 0.057x$
- c. $0.057$, as the number of words increases by 1, we predict the number of sentences increases by 0.057
Problem 9. A researcher wanted to see if she could predict the number of sentences in an advertisement based on the number of words. She collected data on this using a sample of size 54. Below, you are given both a scatterplot of the data as well as the regression output.

![Scatterplot](image)

<table>
<thead>
<tr>
<th>Parameter estimates:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Slope</td>
</tr>
</tbody>
</table>

a. What kind of relationship do you notice between number of words and number of sentences?
b. What is the least squares regression line?
c. What is the estimate for the slope parameter? How would we interpret this?
d. What is the estimate for the intercept parameter? How would we interpret this?
e. The value for r-squared is 0.57. How do we interpret this value? What is the value of the correlation coefficient?
f. What is the 95% confidence interval for the slope parameter?
g. Interpret your interval from part f.

d. 5.39, we predict an advertisement with 8 words will have 5.39 sentences

e. \( r^2 = 0.57, 57\% \) of the variation in # of sentences is explained by the model (# of words)

\[
r = \sqrt{0.57} = \pm 0.755
\]
Problem 9. A researcher wanted to see if she could predict the number of sentences in an advertisement based on the number of words. She collected data on this using a sample of size 54. Below, you are given both a scatterplot of the data as well as the regression output.

a. What kind of relationship do you notice between number of words and number of sentences?
b. What is the least squares regression line?
c. What is the estimate for the slope parameter? How would we interpret this?
d. What is the estimate for the intercept parameter? How would we interpret this?
e. The value for \( r^2 \) is 0.57. How do we interpret this value? What is the value of the correlation coefficient?
f. What is the 95% confidence interval for the slope parameter?
g. Interpret your interval from part f.
Problem 9. A researcher wanted to see if she could predict the number of sentences in an advertisement based on the number of words. She collected data on this using a sample of size 54. Below, you are given both a scatterplot of the data as well as the regression output.

![Scatterplot and regression output](image)

<table>
<thead>
<tr>
<th>Parameter estimates:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Slope</td>
</tr>
</tbody>
</table>

a. What kind of relationship do you notice between number of words and number of sentences?
b. What is the least squares regression line?
c. What is the estimate for the slope parameter? How would we interpret this?
d. What is the estimate for the intercept parameter? How would we interpret this?
e. The value for r-squared is 0.57. How do we interpret this value? What is the value of the correlation coefficient?
f. What is the 95% confidence interval for the slope parameter?
g. Interpret your interval from part f.

9. We are 95% confident that the true slope parameter for this relationship is between 0.0431 and 0.0709.
Problem 10. A researcher wanted to see if she could predict the number of three or more syllable words in an advertisement based on the number of words. She collected data on this using a sample of size 54. Below, you are given the regression output. The researcher would like to test to see if there is a positive relationship between these two variables at the 0.01 level.

![Regression output](image)

a. What kind of relationship do you notice between number of words and number of sentences?
b. What is the least squares regression line?
c. What are the hypotheses?
d. What is the significance level?
e. What is the value of the test statistic?
f. What is the p-value?
g. What is the correct decision?
h. What is the appropriate conclusion/interpretation?
i. How would our test have changed if we were looking for any linear relationship, instead of a positive relationship specifically.
j. What is the 98% confidence interval for the slope parameter?

\[ \text{slope} = +, \text{relationship} = \text{positive} \]

\[ \hat{y} = -3.13 + 1.42x \]

\[ H_0 : \beta_1 = \theta \]
\[ H_A : \beta_1 > \theta \]

\[ \alpha = 0.01 \]
Problem 10. A researcher wanted to see if she could predict the number of three or more syllable words in an advertisement based on the number of words. She collected data on this using a sample of size 54. Below, you are given the regression output. The researcher would like to test to see if there is a positive relationship between these two variables at the 0.01 level.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.1255955</td>
<td>3.0195397</td>
</tr>
<tr>
<td>Slope</td>
<td>1.4199436</td>
<td>0.2256326</td>
</tr>
</tbody>
</table>

a. What kind of relationship do you notice between number of words and number of sentences?
b. What is the least squares regression line?
c. What are the hypotheses?
d. What is the significance level?
e. What is the value of the test statistic?
f. What is the p-value?
g. What is the correct decision?
h. What is the appropriate conclusion/interpretation?
i. How would our test have changed if we were looking for any linear relationship, instead of a positive relationship specifically.
j. What is the 98% confidence interval for the slope parameter?

\[ T_S = \frac{\text{point est.} - \text{null value}}{\text{std. error}} \]

\[ = \frac{1.42 - 0}{0.2256} = \frac{1.42}{0.2256} = 6.29 \]

The test statistic is 6.29.

\[ T_S > 3.496 \]

With 52 degrees of freedom, the p-value is less than 0.0005.
Problem 10. A researcher wanted to see if she could predict the number of three or more syllable words in an advertisement based on the number of words. She collected data on this using a sample of size 54. Below, you are given the regression output. The researcher would like to test to see if there is a positive relationship between these two variables at the 0.01 level.

<table>
<thead>
<tr>
<th>Parameter estimates:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter</td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>3.125955</td>
<td>3.0195397</td>
</tr>
<tr>
<td>Slope</td>
<td>1.4199436</td>
<td>0.22563264</td>
</tr>
</tbody>
</table>

a. What kind of relationship do you notice between number of words and number of sentences?
b. What is the least squares regression line?
c. What are the hypotheses?
d. What is the significance level?
e. What is the value of the test statistic?
f. What is the p-value?
g. What is the correct decision?
h. What is the appropriate conclusion/interpretation?
i. How would our test have changed if we were looking for any linear relationship, instead of a positive relationship specifically.
j. What is the 98% confidence interval for the slope parameter?

9. \( p\text{-value} < 0.0005 < 0.01 \)
   \( p\text{-value} < 0.01 \)
   \( p\text{-value} < \alpha \)

h. The data does provide statistically significant evidence of a positive linear relationship between number of words and number of 3 syllable words.
Problem 10. A researcher wanted to see if she could predict the number of three or more syllable words in an advertisement based on the number of words. She collected data on this using a sample of size 54. Below, you are given the regression output. The researcher would like to test to see if there is a positive relationship between these two variables at the 0.01 level.

![Regression Output]

a. What kind of relationship do you notice between number of words and number of sentences?
b. What is the least squares regression line?
c. What are the hypotheses?
d. What is the significance level?
e. What is the value of the test statistic?
f. What is the p-value?
g. What is the correct decision?
h. What is the appropriate conclusion/interpretation?
i. How would our test have changed if we were looking for any linear relationship, instead of a positive relationship specifically.
j. What is the 98% confidence interval for the slope parameter?

- **Two-sided Test**
  - \( H_0: \beta_1 = \phi \)
  - \( p\text{-value} < 0.001 \)

- \( b_1 \pm t^* (SE) \)
  - \( 1.42 \pm (2.403)(0.2256) \)
  - \( 1.42 \pm 0.54 \)

\[ b_1 = 1.42 - 0.54 = 0.88 \]
\[ b_1 = 1.42 + 0.54 = 1.96 \]

\[ 98\% \text{ CI: } (0.88, 1.96) \]