**Week #13: More Relationships and Associations Between Variables**

**Problem 1.** Use the following ANOVA output to answer the questions below. Assume we are conducting this test at the 10% level.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>2</td>
<td>435.59259</td>
<td>217.7963</td>
<td>1.921042</td>
<td>0.1569</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>51</td>
<td>5781.8889</td>
<td>113.37037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>6217.4815</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Based on the type of test that was run, what kind of variables are we studying?
b. How many groups were compared?
c. How many total observations were there?
d. What is the SSG? What is the SSE? What is the SST?
e. What is the MSG? What is the MSE?
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?

**A. One categorical variable & one numerical variable**

b. \( df_G = K - 1 \)

\[ K = df_G + 1 \]

\[ K = 2 + 1 \]

\[ K = 3 \]

c. \( df_T = n - 1 \)

\[ n = df_T + 1 \]

\[ n = 53 + 1 = 54 \]
**Week #13: More Relationships and Associations Between Variables**

**Problem 1.** Use the following ANOVA output to answer the questions below. Assume we are conducting this test at the 10% level.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>2</td>
<td>435.59259</td>
<td>217.7963</td>
<td>1.9211042</td>
<td>0.1569</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>51</td>
<td>5781.8889</td>
<td>113.37037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>6217.4815</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Based on the type of test that was run, what kind of variables are we studying?
b. How many groups were compared?
c. How many total observations were there?
d. What is the SSG? What is the SSE? What is the SST?
e. What is the MSG? What is the MSE?
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?

\[
\begin{align*}
\text{d. } & \text{SSG} = 435.60 \\
& \text{SSE} = 5781.89 \\
& \text{SST} = 6217.48
\end{align*}
\]

\[
\begin{align*}
\text{e. } & \text{MSG} = 217.80 \\
& \text{MSE} = 113.37
\end{align*}
\]

\[
\begin{align*}
\text{f. } & 1.92
\end{align*}
\]
Problem 1. Use the following ANOVA output to answer the questions below. Assume we are conducting this test at the 10% level.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>2</td>
<td>435.59259</td>
<td>217.7963</td>
<td>1.9211042</td>
<td>0.1569</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>51</td>
<td>5781.8889</td>
<td>113.37037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>6217.4815</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Based on the type of test that was run, what kind of variables are we studying?
b. How many groups were compared?
c. How many total observations were there?
d. What is the SSG? What is the SSE? What is the SST?
e. What is the MSG? What is the MSE?
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?

9. \(0.1569\)

h. \(0.1569 > 0.10\)

\[p\text{-value} > \alpha\]

Fail to Reject \(H_0\)

i. The data does not provide statistically significant evidence that there is an association between these two variables.
**Problem 2.** Use the following data set, create and compute the values for the ANOVA table.

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>16</td>
</tr>
</tbody>
</table>

**ANOVA Table**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorority (Groups)</td>
<td>2</td>
<td>0.000193</td>
<td></td>
<td></td>
<td>0.000193</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
df_G = k-1 = 3-1 = 2
\]

\[
df_T = n-1 = 8-1 = 7
\]

\[
df_E = df_T - df_G = 7 - 2 = 5
\]
Problem 2. Use the following data set, create and compute the values for the ANOVA table.

\[
\bar{X}_1 = 10.07 \\
\bar{X}_2 = 7.5 \\
\bar{X}_3 = 15
\]

\[
\bar{X}_{\text{Grand}} = 11.875
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorority (Groups)</td>
<td>2</td>
<td>93.69</td>
<td>46.845</td>
<td>4.36</td>
<td>0.00193</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
SSG = \sum n_i (X_i - \bar{X}_{\text{Grand}})^2
\]

\[
= 3(10.07 - 11.875)^2 + 2(7.5 - 11.875)^2 + 3(15 - 11.875)^2
\]

\[
= 4.36 + 38.28 + 51.05
\]

\[
= 93.69
\]
Problem 2. Use the following data set, create and compute the values for the ANOVA table.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>16</td>
</tr>
</tbody>
</table>

\[ \overline{X}_1 = 10.07 \]
\[ \overline{X}_2 = 7.5 \]
\[ \overline{X}_3 = 10 \]

\[ \overline{X}_{\text{Grand}} = 11.875 \]

ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorority (Groups)</td>
<td>2</td>
<td>93.69</td>
<td></td>
<td></td>
<td>0.000193</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>5</td>
<td>96.878</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>190.878</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ SST = \sum (X_i - \overline{X}_{\text{Grand}})^2 \]
\[ = (11 - 11.875)^2 + (10 - 11.875)^2 + (11 - 11.875)^2 + (7 - 11.875)^2 + (8 - 11.875)^2 + (15 - 11.875)^2 + (17 - 11.875)^2 + (16 - 11.875)^2 \]
\[ = 0.766 + 3.516 + 0.766 + 23.766 + 18.016 + 9.766 + 26.266 + 17.016 \]
\[ = 96.878 \]
Problem 2. Use the following data set, create and compute the values for the ANOVA table.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorority (Groups)</td>
<td>2</td>
<td>93.109</td>
<td></td>
<td></td>
<td>0.000193 p-value</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>5</td>
<td>3.188</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>96.878</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\overline{x}_1 = 10.07 \\
\overline{x}_2 = 7.5 \\
\overline{x}_3 = 10 \\
\]

\[
\overline{x}_{\text{Grand}} = 11.875
\]

\[
\text{SSE} = \text{SST} - \text{SSG} \\
= 96.878 - 93.109 \\
= 3.188
\]
**Problem 2.** Use the following data set, create and compute the values for the ANOVA table.

\[
\begin{align*}
\bar{x}_1 &= 10.07 \\
\bar{x}_2 &= 7.5 \\
\bar{x}_3 &= 10
\end{align*}
\]

\[
\bar{x}_{\text{Grand}} = 11.875
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sorority (Groups)</strong></td>
<td>2</td>
<td>93.09</td>
<td>46.845</td>
<td>46.845</td>
<td>0.000193</td>
</tr>
<tr>
<td><strong>Error (Residuals)</strong></td>
<td>5</td>
<td>3.188</td>
<td>0.634</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>7</td>
<td>96.878</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{MSG} = \frac{SSG}{df_G} = \frac{93.09}{2} = 46.845
\]

\[
\text{MSE} = \frac{SSE}{df_E} = \frac{3.188}{5} = 0.634
\]
Problem 2. Use the following data set, create and compute the values for the ANOVA table.

\[ \bar{x}_1 = 10.07 \]
\[ \bar{x}_2 = 7.5 \]
\[ \bar{x}_3 = 10 \]

ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorority (Groups)</td>
<td>2</td>
<td>93.169</td>
<td>46.584</td>
<td>73.89</td>
<td>0.000193</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>5</td>
<td>3.188</td>
<td>0.636</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>96.878</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ F = \frac{\text{MSG}}{\text{MSE}} = \frac{46.584}{0.634} = 73.89 \]
**Problem 3.** A teacher wants to know if test scores are different between her classes. Test this at the 5% level.

a. What kind of test should we conduct?
b. What are the hypotheses?
c. What is the significance level?
d. The ANOVA table below is partially filled in. Complete the missing spaces.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class (Groups)</td>
<td>3</td>
<td>3.798</td>
<td>0.066358</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td></td>
<td>41.228</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>41.228</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e. How many different classes was the teacher comparing?
f. How many total students were in those classes?
g. What is the value of the test statistic?
h. What is the p-value?
i. What is the correct decision?
j. What is the appropriate conclusion/interpretation?

**a. ANOVA**

**b.** \( H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \)

**HA:** At least one is different

**c.** \( \alpha = 0.05 \)
Problem 3. A teacher wants to know if test scores are different between her classes. Test this at the 5% level.

a. What kind of test should we conduct?

b. What are the hypotheses?

c. What is the significance level?

d. The ANOVA table below is partially filled in. Complete the missing spaces.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class (Groups)</td>
<td>3</td>
<td>3.798</td>
<td>0.066358</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>81</td>
<td>41.228</td>
<td></td>
<td></td>
<td>0.066358</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td></td>
<td>45.026</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e. How many different classes was the teacher comparing?

f. How many total students were in those classes?

g. What is the value of the test statistic?

h. What is the p-value?

i. What is the correct decision?

j. What is the appropriate conclusion/interpretation?

d. \( df_E = df_T - df_E = 84 - 3 = 81 \)

\[ \text{SST} = \text{SSG} + \text{SSE} \]

\[ = 3.798 + 41.228 \]

\[ = 45.026 \]
Problem 3. A teacher wants to know if test scores are different between her classes. Test this at the 5% level.

a. What kind of test should we conduct?
b. What are the hypotheses?
c. What is the significance level?
d. The ANOVA table below is partially filled in. Complete the missing spaces.

e. How many different classes was the teacher comparing?
f. How many total students were in those classes?
g. What is the value of the test statistic?
h. What is the p-value?
i. What is the correct decision?
j. What is the appropriate conclusion/interpretation?

\[
\text{d. } MSG = \frac{SSG}{df_G} = \frac{3.798}{3} = 1.266
\]

\[
\text{MSE} = \frac{SSE}{df_E} = \frac{41.228}{81} = 0.509
\]

\[
F = \frac{MSG}{MSE} = \frac{1.266}{0.509} = 2.487
\]
**Problem 3.** A teacher wants to know if test scores are different between her classes. Test this at the 5% level.

a. What kind of test should we conduct?

b. What are the hypotheses?

c. What is the significance level?

d. The ANOVA table below is partially filled in. Complete the missing spaces.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class (Groups)</td>
<td>3</td>
<td>3.798</td>
<td>1.2666</td>
<td>2.487</td>
<td>0.066358</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>81</td>
<td>41.228</td>
<td>0.509</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>45.026</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e. How many different classes was the teacher comparing?

f. How many total students were in those classes?

g. What is the value of the test statistic?

h. What is the p-value?

i. What is the correct decision?

j. What is the appropriate conclusion/interpretation?

e. $df_G = k - 1$

   $k = 4$

f. $df_T = n - 1$

   $n = 85$

9. $2.487$

h. $0.066358 > 0.05$

   p-value > $\alpha$

   Fail to Reject

i. The data does not provide statistically significant evidence that there is an association between class and test score.
Problem 4. A commuter thinks that the time it takes them to commute is different based on what day of the week it is. For a few weeks, they record their commute time each day, Monday through Friday. Test their claim at the 5% level.

a. What kind of test should we conduct?
b. What are the hypotheses?
c. What is the significance level?
d. The ANOVA table below is partially filled in. Complete the missing spaces.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day (Groups)</td>
<td></td>
<td>14.28</td>
<td>0.037122</td>
<td>0.037122</td>
<td></td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>30.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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?

a. ANOVA

b. $H_0: \mu_M = \mu_{Tu} = \mu_W = \mu_{Th} = \mu_F$

$c. H_A: \text{At least one mean is different}$

d. $\alpha = 0.05$
Problem 4. A commuter thinks that the time it takes them to commute is different based on what day of the week it is. For a few weeks, they record their commute time each day, Monday through Friday. Test their claim at the 5% level.

a. What kind of test should we conduct?

b. What are the hypotheses?

c. What is the significance level?

d. The ANOVA table below is partially filled in. Complete the missing spaces.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day (Groups)</td>
<td>4</td>
<td>14.28</td>
<td>3.57</td>
<td>3.37</td>
<td>0.037122</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>15</td>
<td>15.92</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>30.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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?

d. \( \text{df}_G = K-1 = 5-1 = 4 \)

\( \text{df}_E = \text{df}_T - \text{df}_G = 19-4 = 15 \)

\( \text{SSE} = \text{SST} - \text{SSG} = 30.2 - 14.28 = 15.92 \)

\( \text{MSG} = \frac{\text{SSG}}{\text{df}_G} = \frac{14.28}{4} = 3.57 \)

\( \text{MSE} = \frac{\text{SSE}}{\text{df}_E} = \frac{15.92}{15} = 1.06 \)

\( F = \frac{\text{MSG}}{\text{MSE}} = \frac{3.57}{1.06} = 3.37 \)
Problem 4. A commuter thinks that the time it takes them to commute is different based on what day of the week it is. For a few weeks, they record their commute time each day, Monday through Friday. Test their claim at the 5% level.

a. What kind of test should we conduct?
b. What are the hypotheses?
c. What is the significance level?
d. The ANOVA table below is partially filled in. Complete the missing spaces.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day (Groups)</td>
<td>4</td>
<td>14.28</td>
<td>3.57</td>
<td>3.37</td>
<td>0.037122</td>
</tr>
<tr>
<td>Error (Residuals)</td>
<td>15</td>
<td>15.92</td>
<td>1.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>30.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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?

e. 3.37
f. 0.037
g. 0.037 < 0.05
p-value < \( \alpha \)
Reject \( H_0 \)

h. The data does provide statistically significant evidence that commute time and day of the week are associated.
Problem 5. 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%.

\[
\begin{align*}
X &= 30.8', \\
\bar{X} &= 26.7', \\
Y &= 38.8', \\
\bar{Y} &= 16.9'.
\end{align*}
\]

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.

\[
\begin{align*}
a. & \quad r^2 = 0.72 \\
& \quad r = \sqrt{0.72} \\
& \quad r = \pm 0.849 \\
\text{based on scatterplot: } & \quad r = -0.849
\end{align*}
\]
Problem 5. 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.
Problem 5. 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 reduced-fee lunch, we predict 55.34% of bike riders 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 5. 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%.

![Scatterplot showing the relationship between percentage of children receiving reduced-fee lunches and percentage of bike riders wearing helmets.](image)

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.

\[
\hat{y} = 55.34 - 0.537x
\]

\[
\hat{y} = 55.34 - 0.537(40)
\]

\[
\hat{y} = 33.86
\]

\[
\text{Residual} = y - \hat{y} = 40 - 33.86 = 6.14
\]

- A positive value, so the actual value is higher than the predicted value. Actual value overestimates prediction.
Problem 6. 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.

- Does $\beta_1 = 0$ or not?

| 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 |

\[ df = n - 2 \]

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. negative relationship  
  weak-moderate relationship

b. These points are high leverage points—potential to be influential  
  - We are leaving out data?

C. $\hat{y} = -7.3644 - 0.8897x$
a. $H_0: \beta_1 = \emptyset$
   $H_A: \beta_1 \neq \emptyset$

e. $\alpha = 0.10$

f. $T_S = \frac{\text{point estimate} - \text{null value}}{\text{standard error}}$

$= \frac{-0.8897 - \emptyset}{0.8350} = \frac{-0.8897}{0.8350} = -1.0600$
9. $TS = t = -1.060$
   \[ df = 27 \]
   Two-sided test ($H_A: \neq$)
   \[ 1.087 < |TS| < 1.314 \]
   \[ 0.20 < p-value < 0.30 \]

10. $p-value > 0.20 > 0.10$
    $p-value > 0.10$
    $p-value > \alpha$
    Fail to Reject

i. The data does not provide statistically significant evidence that there is a linear relationship between unemployment and midterm election results.
\[ b_1 = t \times \text{std. error} \]

\[-0.8897 \pm (2.052)(0.8350)\]

\[-0.8897 \pm 1.71342\]

95\% CI: (-2.00, 0.82)
Problem 7. 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 showing the relationship between number of words and number of sentences]

<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.

a. positive, linear, moderate

b. \( \hat{y} = 5.39 + 0.057x \)

c. 0.057, as number of words increases by 1, we predict number of sentences increases by 0.057

d. 5.39, we predict an advertisement with 5 words has 5.39 sentences
Problem 7. 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-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.

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

\[
r = \sqrt{0.57}
\]

\[
r = \pm 0.755
\]

- D scatterplot \[ r = \pm 0.755 \]
Problem 7. 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)

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.

\[
b_1 = 0.057 \\ \text{SE} = 0.0069 \\ t^* : df = n-2 = 52 \]

\[
0.057 \pm 2.009(0.0069) 
\]

\[
0.057 \pm 0.0139 \rightarrow 95\% \text{ CI: (0.0431, 0.0709)}
\]

We are 95% confident that the true slope parameter is between 0.0431 and 0.0709.
Problem 8. 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>Parameter</td>
<td>Estimate</td>
<td>Std. Err.</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?

\[ \text{a. Slope} = + \text{ / relationship } = \text{ positive} \]
\[ \text{b. } \hat{y} = -3.13 + 1.42x \]
\[ \text{c. } H_0: \beta_1 = 0 \]
\[ H_A: \beta_1 > 0 \]
\[ \text{d. } \alpha = 0.01 \]
Problem 8. 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.

![Parameter estimates image](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?

E. \[ TS = \frac{\text{point estimate} - \text{null value}}{\text{std. error}} \]

\[ E. \quad TS = \frac{1.42 - 0}{0.2256} = \frac{1.42}{0.2256} = 6.29 \]

f. \[ df = n - 2 = 54 - 2 = 52 \]

One-sided (\( H_A: > \)) \[ TS > 3.4916 \] \[ p-value < 0.0005 \]
Problem 8. 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>Parameter</td>
<td>Estimate</td>
<td>Std. Err.</td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.1255955</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. \( p\text{-value} < 0.0005 < 0.01 \)
   \[ p\text{-value} < 0.01 \]
   \[ p\text{-value} < \alpha \]
   \[ \text{Rej. } H_0 \]

10. The data does provide statistically significant evidence of a positive linear relationship between \# of words and \# of syllables.
**Problem 8.** 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>
</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 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?

\[ 1.42 \pm t^* (SE) \]
\[ 1.42 \pm (2.403)(0.2256) \]
\[ 1.42 \pm 0.54 \]
\[ 98\% \ CI : (0.88, 1.96) \]