For the ordered data sets \(x = (0, 6, 12, 18)\) and \(y = (3, 3, 5, 9)\), we have \(\bar{x} = 9\) and \(\bar{y} = 5\),
\[
\begin{align*}
s_x^2 &= \frac{1}{3} \left[ (0 - 9)^2 + (6 - 9)^2 + (12 - 9)^2 + (18 - 9)^2 \right] = 60, \\
s_y^2 &= \frac{1}{3} \left[ (3 - 5)^2 + (3 - 5)^2 + (5 - 5)^2 + (9 - 5)^2 \right] = 8, \\
s_{xy} &= \frac{1}{3} \left[ (0 - 9)(3 - 5) + (6 - 9)(3 - 5) + (12 - 9)(5 - 5) + (18 - 9)(9 - 5) \right] = 20.
\end{align*}
\]

For the data sets \(x + y = (3, 9, 17, 27)\) and \(x - y = (-3, 3, 7, 9)\), we have \(x + y = 14\), \(x - y = 4\), and
\[
\begin{align*}
s_{x+y}^2 &= \frac{1}{3} \left[ (3 - 14)^2 + (9 - 14)^2 + (17 - 14)^2 + (27 - 14)^2 \right] = 108, \\
s_{x-y}^2 &= \frac{1}{3} \left[ (-3 - 4)^2 + (3 - 4)^2 + (7 - 4)^2 + (9 - 4)^2 \right] = 28.
\end{align*}
\]

From these calculations, it follows that
\[
\begin{align*}
\sqrt{x+y} &= 14 = 9 + 5 = \bar{x} + \bar{y} \quad \checkmark \\
\sqrt{s_{x+y}^2} &= 108 = 60 + 8 + 2(20) = s_x^2 + s_y^2 + 2s_{xy} \quad \checkmark \\
\sqrt{x-y} &= 4 = 9 - 5 = \bar{x} - \bar{y} \quad \checkmark \\
\sqrt{s_{x-y}^2} &= 28 = 60 + 8 - 2(20) = s_x^2 + s_y^2 - 2s_{xy} \quad \checkmark
\end{align*}
\]
For the data sets \( x = (0, 6, 12, 18) \) and \( y = (3, 9, 3, 5) \), we have \( \overline{x} = 9 \), \( \overline{y} = 5 \),

\[
\begin{align*}
s_x^2 &= \frac{1}{3} \left[ (0 - 9)^2 + (6 - 9)^2 + (12 - 9)^2 + (18 - 9)^2 \right] = 60, \\
s_y^2 &= \frac{1}{3} \left[ (3 - 5)^2 + (9 - 5)^2 + (3 - 5)^2 + (5 - 5)^2 \right] = 8, \\
s_{xy} &= \frac{1}{3} \left[ (0 - 9)(3 - 5) + (6 - 9)(9 - 5) + (12 - 9)(3 - 5) + (18 - 9)(5 - 5) \right] = 0. \\
\end{align*}
\]

For the data sets \( x + y = (3, 15, 15, 23) \) and \( x - y = (-3, -3, 9, 13) \), we have \( \overline{x+y} = 14 \), \( \overline{x-y} = 4 \) as before, but

\[
\begin{align*}
s_{x+y}^2 &= \frac{1}{3} \left[ (3 - 14)^2 + (15 - 14)^2 + (15 - 14)^2 + (23 - 14)^2 \right] = 68, \\
s_{x-y}^2 &= \frac{1}{3} \left[ (-3 - 4)^2 + (-3 - 4)^2 + (9 - 4)^2 + (13 - 4)^2 \right] = 68. \\
\end{align*}
\]

From these calculations, it follows that

\[
\begin{align*}
\overline{x+y} &= 14 = 9 + 5 = \overline{x} + \overline{y} & \checkmark \\
\frac{1}{3} (x+y)^2 &= 68 = 60 + 8 + 2(0) = s_x^2 + s_y^2 + 2s_{xy} & \checkmark \\
\overline{x-y} &= 4 = 9 - 5 = \overline{x} - \overline{y} & \checkmark \\
\frac{1}{3} (x-y)^2 &= 68 = 60 + 8 - 2(0) = s_x^2 + s_y^2 - 2s_{xy}. & \checkmark \\
\end{align*}
\]

* Note that, since the datasets \( x \) and \( y \) are paired, they cannot be independent, even though their sample covariance = 0. However, if two random variables \( X \) and \( Y \) are independent, then their population covariance = 0, and this case might be suggested by two sample datasets \( x \) and \( y \) having a zero (or small) covariance. Check the scatterplot!
The mean time is $\bar{x} = 6$ months, and the mean assay is $\bar{y} = 448$ mg. Hence the sample variances and sample covariance are, respectively,

$$s_x^2 = \frac{(0 - 6)^2 + (3 - 6)^2 + (6 - 6)^2 + (9 - 6)^2 + (12 - 6)^2}{5 - 1} = \frac{(-6)^2 + (-3)^2 + 0^2 + 3^2 + 6^2}{4} = 22.5$$

$$s_y^2 = \frac{(500 - 448)^2 + (490 - 448)^2 + (470 - 448)^2 + (430 - 448)^2 + (350 - 448)^2}{5 - 1}$$

$$= \frac{52^2 + 42^2 + 22^2 + (-18)^2 + (-98)^2}{4} = \frac{14880}{4} = 3720$$

$$s_{xy} = \frac{(-6)(52) + (-3)(42) + (0)(22) + (3)(-18) + (6)(-98)}{5 - 1} = -270$$

Thus, $r = \frac{s_{xy}}{s_x s_y} = \frac{-270}{\sqrt{22.5} \sqrt{3720}} = -0.933$, indicating a strong negative linear correlation.
(b) \( b_1 = \frac{-270}{22.5} = -12, \quad b_0 = 448 - (-12)(6) = 520 \)

Hence the equation of the least squares regression line is \( \hat{Y} = 520 - 12X \).

The 95% confidence limits for the slope \( \beta_1 \) are \( b_1 \pm t_{0.025, n-2} \cdot s_e \frac{1}{\sqrt{S_{xx}}} \), where \( s_e^2 = \text{MSError} \)

\[
\frac{SS_{\text{Error}}}{n-2} = \frac{1920}{3} = 640, \quad \text{and} \quad S_{xx} = (n-1) s_x^2 = (4)(22.5) = 90. \quad \text{Therefore, in this case, we have the limits} \\
-12 \pm t_{0.025, 3} \cdot \left( \frac{8}{3} \right) = -12 \pm (3.182)(\frac{8}{3}) = -12 \pm 8.485. \quad \text{So the 95% confidence interval for the slope is} \ (-20.485, -3.515). \\
\]

(c)

<table>
<thead>
<tr>
<th>predictor X</th>
<th>0</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed Y</td>
<td>500</td>
<td>490</td>
<td>470</td>
<td>430</td>
<td>350</td>
</tr>
<tr>
<td>fitted ( \hat{Y} )</td>
<td>520</td>
<td>484</td>
<td>448</td>
<td>412</td>
<td>376</td>
</tr>
<tr>
<td>( Y - \bar{Y} )</td>
<td>+52</td>
<td>+42</td>
<td>+22</td>
<td>-18</td>
<td>-98</td>
</tr>
<tr>
<td>( \hat{Y} - \bar{Y} )</td>
<td>+72</td>
<td>+36</td>
<td>0</td>
<td>-36</td>
<td>-72</td>
</tr>
<tr>
<td>( e = Y - \hat{Y} )</td>
<td>-20</td>
<td>+6</td>
<td>+22</td>
<td>+18</td>
<td>-26</td>
</tr>
</tbody>
</table>

\[ \text{SS}_{\text{Total}} = 14880 \quad \text{(calculated above in } s_y^2) \]

\[ \text{SS}_{\text{Reg}} = 12960 \quad \text{SS}_{\text{Error}} = 1920 \]

\[ \text{By construction, this is the smallest value that the “residual sum of squares” can attain, of any regression line for these data.} \]

(d) ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>12960</td>
<td>12960</td>
<td>20.25</td>
<td>.01 &lt; p &lt; .05</td>
</tr>
<tr>
<td>Error</td>
<td>3</td>
<td>1920</td>
<td>640</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>14880</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(e) \( r^2 = (-0.933)^2 = 0.871 \quad \text{Method 2:} \quad r^2 = \frac{12960}{14880} = 0.871 \)

**Interpretation:** The regression line accounts for 87.1% of the variation in the sample data; the remaining 12.9% is unaccounted for by the linear model, and may be due to random variation. (This DOES NOT mean that 87.1% of the sample data points lie on the line. This is clearly false; in fact, none of the data points actually lies on the regression line!)
(f) To test $H_0: \rho = 0$ versus $H_A: \rho \neq 0$, we can calculate the $p$-value = $2 P\left( T_3 \leq -\frac{0.933\sqrt{3}}{\sqrt{1 - (-0.933)^2}} \right) = 2 P(T_3 \leq -4.5) \approx .02 < .05$, the significance level. From this, we can reject the null hypothesis, and conclude that there is indeed evidence of a linear association between $X$ and $Y$. We can reach the same conclusion by rejecting the null hypothesis $H_0: \beta_1 = 0$ in favor of the alternative $H_A: \beta_1 \neq 0$, via observing that $.01 < p < .05$ from the ANOVA table*, and from the fact that the confidence interval $(-20.485, -3.515)$ does not contain $\beta_1 = 0$.

* Note that this is consistent with the previous finding. In fact, the two tests are equivalent; the square of the $t$-score is equal to the $F$-ratio, i.e., $(−4.5)^2 = 20.25$.

(g) When $X = 6$ months, it follows that $\hat{Y} = 520 - 12 (6) = 448$ mg is the estimated potency. [Note that this is the “center of mass” $(\bar{x}, \bar{y})$, which always lies on the least squares regression line.] However, it is not particularly realistic, for this value should occur between 6 and 9 months, judging from the data. The corresponding 95% confidence limits are $(b_0 + b_1 x^*) \pm t_{0.025}, n - 2 \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = 448 \pm (3.182) \sqrt{640 \cdot \frac{1}{5} + \frac{(6 - 6)^2}{90}} = 448 \pm 36$. Therefore, the 95% confidence interval for the mean potency at 6 months is $(412, 484)$ mg which, while indeed including the observed potency of 470 mg, is nevertheless quite large.

(h) Ninety percent of the labeled 500 mg potency is equal to 450 mg. Thus, solving for $X$ in the linear equation $450 = 520 - 12 X$ yields the expiration date $X = 5.833$ months. Again, this is not especially realistic, for the potency at this time is approximately 470 mg, judging from the data; 450 mg seems to occur between 6 and 9 months. These observations indicate that it may be possible to improve the model, despite the high $r^2$ value.

(i)

![Residuals vs Fitted](image)

If $\hat{X} = X / 3$ and $\hat{Y} = 510 - Y$, then the transformed data become

$$
\begin{array}{c|c|c|c|c|c}
X & 0 & 1 & 2 & 3 & 4 \\
\hline
\hat{X} & 0 & 1 & 2 & 3 & 4 \\
\hat{Y} & 10 & 20 & 40 & 80 & 160 \\
\end{array}
$$

where $\hat{Y}$ doubles with every unit increase in $\hat{X}$. Thus, if we regress on a logarithmic response scale, i.e., $V = \ln(\hat{Y})$ vs. $\hat{X}$, then the linear model

$$
V = 2.303 + 0.693 \hat{X}
$$

is a better (in fact, exact) fit. Backsolving yields $\hat{Y} = 10 \cdot 2^{\hat{X}/3}$ as expected, i.e.,

$$
V = 510 - 10 \cdot 2^{\hat{X}/3}
$$

When $Y = 450$ mg, we obtain $X = \frac{3 \ln 6}{\ln 2} = 7.75$ months, which is much more consistent.

(j) Straightforward; $p$-value occurs once for $t$-test, and once for ANOVA overall $F$-test, of $H_0: \beta_1 = 0$. 

The reciprocal transformation \( Y = \frac{1}{\alpha X + \beta} \rightarrow \frac{1}{Y} = \alpha X + \beta \) converts the given nonlinear data into the linear plot shown above, containing the points \((0, \frac{1}{60}), (1, \frac{1}{30}), \ldots, (5, \frac{1}{10})\). Because these points are actually collinear, it is algebraically possible to solve for the linear equation exactly, without the need for regression formulas. The slope of this line is \( \alpha = \frac{1}{60} \); the vertical intercept is \( \beta = \frac{1}{60} \) as well. Therefore, the relation is given by \( \frac{1}{Y} = \frac{1}{60} X + \frac{1}{60} \), i.e., \( Y = \frac{60}{X+1} \), which matches the original data exactly. (Check it!)
7-4. Equality of Test Statistics

(a) The third formula \( r^2 = 1 - \frac{SS_{Err}}{SS_{Tot}} \) can be rewritten as \( SS_{Err} = (1 - r^2)SS_{Tot} \); substituting the fourth formula \( SS_{Tot} = (n - 1)s_y^2 \) into this yields \( SS_{Err} = (1 - r^2)(n - 1)s_y^2 \). Therefore, the second formula becomes \( MS_{Err} = \frac{(1 - r^2)(n - 1)s_y^2}{n - 2} \). So...

\[
T\text{-score} = \left(\frac{b_1 - \beta_1}{\sqrt{MS_{Err}}}\right)\sqrt{S_{xx}} = \frac{\left(\frac{s_{xy}}{s_x^2} - 0\right)}{\sqrt{(1 - r^2)(n - 1)s_y^2}} \sqrt{(n - 1)s_x^2} = \frac{\left(\frac{s_{xy}}{s_x s_y}\right)}{\frac{1 - r^2}{\sqrt{n - 2}}} = r \sqrt{\frac{n - 2}{1 - r^2}}. \quad \checkmark
\]

(b) \( (T\text{-score})^2 = \left(\frac{b_1 - \beta_1}{\sqrt{MS_{Err}}}\right)^2 \left(\sqrt{S_{xx}}\right)^2 = \frac{\left(\frac{s_{xy}}{s_x^2} - 0\right)^2}{MS_{Err}} S_{xx} = \frac{\left(\frac{s_{xy}}{s_x^2}\right)^2}{MS_{Err}} \left(\frac{1 - r^2}{\sqrt{n - 2}}\right) = \frac{(n - 1)s_x^2}{MS_{Err}} \]

\[
\frac{(n - 1)\left(\frac{s_{xy}}{s_x}\right)^2}{MS_{Err}} = \frac{(n - 1)s_y^2 \left(\frac{s_{xy}}{s_x s_y}\right)^2}{MS_{Err}} = SS_{Tot} \frac{r^2}{MS_{Err}} = SS_{Reg} \frac{MS_{Reg}}{MS_{Err}} = MS_{Reg} = F\text{-ratio}. \quad \checkmark
\]

* because \( MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} \), but \( df_{Reg} = 1 \), so that \( MS_{Reg} = SS_{Reg} \).
7-5. Scatterplot, Least Squares Regression Line, and Residuals

\[
\hat{Y} = 194 + 1.5X
\]

(a) \( s_{xy} = \frac{1}{5-1} \left[ (0-4)(200-200) + (2-4)(190-200) + (4-4)(210-200) + (6-4)(180-200) + (8-4)(220-200) \right] \)

\[
= \frac{1}{4} \left[ (-4)(0) + (-2)(-10) + (0)(10) + (2)(-20) + (4)(20) \right] = \frac{1}{4} (60) = +15
\]

(b) \( r = \frac{+15}{\sqrt{10} \sqrt{250}} = +0.3 \); a weak positive linear correlation between \( X \) and \( Y \).

(c) \( b_1 = \frac{15}{10} = 1.5, \quad b_0 = 200 - (1.5)(4) = 194 \)

Hence the equation of the least squares regression line is \( \hat{Y} = 194 + 1.5X \).

(d) 

<table>
<thead>
<tr>
<th>predictors</th>
<th>( X )</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed responses</td>
<td>( Y )</td>
<td>200</td>
<td>190</td>
<td>210</td>
<td>180</td>
<td>220</td>
</tr>
<tr>
<td>fitted responses</td>
<td>( \hat{Y} )</td>
<td>194</td>
<td>197</td>
<td>200</td>
<td>203</td>
<td>206</td>
</tr>
<tr>
<td>residuals</td>
<td>( Y - \hat{Y} )</td>
<td>+6</td>
<td>-7</td>
<td>+10</td>
<td>-23</td>
<td>+14</td>
</tr>
</tbody>
</table>

Note that \( SS_{\text{Error}} = (6)^2 + (-7)^2 + (10)^2 + (-23)^2 + (14)^2 = 910 \).

By construction, this is the smallest value that the “residual sum of squares” can attain, of any regression line for these data.

Exercise: Construct the ANOVA table for this regression problem.
(e) \( r^2 = (0.3)^2 = +0.09 \)

Interpretation: Interpreting this quantity as the equivalent ratio \( SS_{\text{Reg}}/SS_{\text{Total}} \), the linear association between month and weight accounts for only 9% of the total variation in the data. (It DOES NOT mean that 9% of the sample data points lie on the line. This is clearly false; in fact, none of the data points actually lies on the regression line!) The remaining 91% of the variation is unaccounted for by the linear regression model; such a whoppingly large fraction is certainly very strong evidence that the model is inadequate.

(f) Interpretation: Patterns of weight change among compulsive overeaters are well documented. Such individuals tend to lose weight at the beginning of their dieting programs, then binge, and regain all the lost weight, plus more. The cycle of weight fluctuation repeats over time, and can lead to frustration, depression, and more severe medical complications. The scatterplot of the data values (as well as the residual plot, below) clearly shows this nonlinear pattern between the variables which, not surprisingly, the regression line fails to capture. In fact, all three formal assumptions of least squares regression are violated here:

- As noted above, this model has a very low correlation coefficient \( r \), and correspondingly low \( r^2 \) value, indicating a generally poor fit. The model is not useful, let alone “correct.”
- The error values are not randomly distributed over time; in particular, they are not independent of one another, for a weight loss (i.e., negative residual) in any 2-month period is followed by a weight gain (i.e., positive residual) in the next 2-month period, and vice versa.
- The error values have nonconstant variance over time; in particular, the amount that weight varies clearly grows larger, as more and more weight is lost and subsequently regained. (Note how the residuals tend to fan out below, instead of remaining bounded.)
7-6. (a) **Scatterplot, Least Squares Regression Line, and Residuals**

![Scatterplot, Least Squares Regression Line, and Residuals](image)

(b) \( r = \frac{260}{\sqrt{250} \sqrt{275}} = +0.992 \); a **strong positive** linear correlation between \( X \) and \( Y \).

(c) \( b_1 = \frac{260}{250} = 1.04, \quad b_0 = 52 - (1.04) (50) = 0 \)

Hence the equation of the least squares regression line is \( Y = 1.04X \).

<table>
<thead>
<tr>
<th>predictors</th>
<th>( X )</th>
<th>( 30.0 )</th>
<th>( 40.0 )</th>
<th>( 50.0 )</th>
<th>( 60.0 )</th>
<th>( 70.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed responses</td>
<td>( Y )</td>
<td>( 32.0 )</td>
<td>( 39.0 )</td>
<td>( 53.0 )</td>
<td>( 65.0 )</td>
<td>( 71.0 )</td>
</tr>
<tr>
<td>fitted responses</td>
<td>( \hat{Y} )</td>
<td>( 31.2 )</td>
<td>( 41.6 )</td>
<td>( 52.0 )</td>
<td>( 62.4 )</td>
<td>( 72.8 )</td>
</tr>
<tr>
<td>residuals</td>
<td>( Y - \hat{Y} )</td>
<td>( +0.8 )</td>
<td>( -2.6 )</td>
<td>( +1.0 )</td>
<td>( +2.6 )</td>
<td>( -1.8 )</td>
</tr>
</tbody>
</table>

Note that \( \text{SS}_{\text{Error}} = (0.8)^2 + (-2.6)^2 + (1.0)^2 + (2.6)^2 + (-1.8)^2 = 18.4 \); this is the smallest value that the “residual sum of squares” can attain, of any regression line for these data.
(d) ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1081.6</td>
<td>1081.6</td>
<td>176.35</td>
<td>&lt;&lt; .05</td>
</tr>
<tr>
<td>Error</td>
<td>3</td>
<td>18.4</td>
<td>6.133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>1100.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that \( \frac{SS_{\text{Total}}}{df_{\text{Total}}} = s^2_y \), i.e., \( \frac{1100}{5 - 1} = 275 \), given.

(e) 95% margin of error for the slope \( \beta_1 = t_{0.025, 3} \sqrt{\frac{6.133}{4(250)}} = 3.182 \times 0.0783 = 0.2492 \)
\[ \therefore 95\% \text{ confidence interval} = (1.04 - 0.25, 1.04 + 0.25) = (0.79, 1.29) \]

(f) According to the ANOVA table, the p-value is exceedingly small, indicating an extremely strong rejection of the null hypothesis \( H_0: \beta_1 = 0 \) at the \( \alpha = .05 \) significance level. This is the same conclusion reached via the 95% confidence interval, which does not contain the null value of 0.

**Interpretation:** There is strong evidence that there exists a significant linear association between the known assay amounts \( X \) and the empirically derived assay amounts \( Y \). Of course, this is not at all surprising; the real question of interest is whether or not the linear association is significantly different from the ideal calibration line \( Y = X \).

(g) The 95% confidence interval does contain the null value \( H_0: \beta_1 = 1 \). Hence this hypothesis is not rejected at the \( \alpha = .05 \) level.

**Interpretation:** Based on this sample of five observations, there is no significant difference between the results of this new assay procedure, and an ideal error-free procedure that exactly agrees with known amounts. If the new procedure is cheaper than an existing one, for example, then it may prove to be a useful alternative.