A. Using Population Regression Lines

• 10.3

(a) $\beta_0 = 0.3$

The mean overseas market return is 0.3% when the U.S. market is flat.

(b) $\beta_1 = 0.85$

For each 1% increase in return in the U.S. market, the mean overseas market increases by 0.85%.

(c) There are two different (but equivalent) ways to write the population regression line. One way emphasizes the mean response $\mu_y$. The other way emphasizes a particular response $y$:

\[
\begin{align*}
\mu_y &= \beta_0 + \beta_1 x \\
y &= \beta_0 + \beta_1 x + \varepsilon
\end{align*}
\]

Looking at these two equations together shows that

\[
y - \mu_y = \varepsilon
\]

so the “error” $\varepsilon = y - \mu_y$ represents the difference between a particular response $y$ and the mean response $\mu_y$.

○ So a year with a positive error $\varepsilon$ means that $y > \mu_y$ (the overseas return for that year is better than expected)

○ A year with a negative error $\varepsilon$ means that $y < \mu_y$ (the overseas return for that year is worse than expected)

To answer the book’s question:

\[
\begin{align*}
\text{Mean Overseas Return} &= 0.3 + 0.85 \times \text{(U.S. Return)} \\
\text{Overseas Return} &= 0.3 + 0.85 \times \text{(U.S. Return)} + \varepsilon
\end{align*}
\]

so

$\varepsilon = \text{Overseas Return} - \text{Mean Overseas Return}$

allows overseas returns to vary in different years, even if U.S. returns remain the same.

• 10.4

<table>
<thead>
<tr>
<th>Total cost</th>
<th>Number of units</th>
</tr>
</thead>
</table>

(a) $y = \beta_0 + \beta_1 x + \varepsilon$

(b) $\beta_0$

(c) $\beta_1$. We expect $\beta_1 > 0$ since cost of production increases as more items are produced.

(d) The “error” $\varepsilon$ allows data points $(x, y)$ to vary from the regression line.

(That is, the cost $y$ can differ, even when producing the same number of units $x$.)
B. Research and Development Spending

• 10.5

(a) There is a strong positive linear relationship between Year and Spending.

(b) Entering data into a calculator and using the calculator’s regression function:

\[ \text{Spending} = -4566.24 + 2.3 \text{ Year} \]

(c)

<table>
<thead>
<tr>
<th>x = Year</th>
<th>y = Spending</th>
<th>$\hat{y}$ = Predicted Spending</th>
<th>$e = y - \hat{y}$</th>
<th>$(y - \hat{y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>40.1</td>
<td>40.66</td>
<td>-0.56</td>
<td>0.3136</td>
</tr>
<tr>
<td>2004</td>
<td>43.3</td>
<td>42.96</td>
<td>0.34</td>
<td>0.1156</td>
</tr>
<tr>
<td>2005</td>
<td>45.8</td>
<td>45.26</td>
<td>0.54</td>
<td>0.2916</td>
</tr>
<tr>
<td>2006</td>
<td>47.7</td>
<td>47.56</td>
<td>0.14</td>
<td>0.0196</td>
</tr>
<tr>
<td>2007</td>
<td>49.4</td>
<td>49.86</td>
<td>-0.46</td>
<td>0.2116</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0.952</td>
</tr>
</tbody>
</table>

\[ s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{0.952}{5-2} = \frac{0.952}{3} = 0.3173 \]

\[ s = \sqrt{s^2} = \sqrt{0.3173} = 0.5633 \]

(d) The population regression model:

\[ y = \beta_0 + \beta_1 \times x + \varepsilon \]

OR \[ \text{Spending} = \beta_0 + \beta_1 \times \text{Year} + \varepsilon \]

\[ \hat{\beta}_0 = b_0 = -4566.24 \]
\[ \hat{\beta}_1 = b_1 = 2.3 \]

(e)

\[ \hat{y} = b_0 + b_1 x \]
\[ = -4566.24 + 2.3x \]
\[ = -4566.24 + (2.3)(2001) = 36.06 \]

So prediction error for 2001 is \[ e = y - \hat{y} = 32.8 - 36.06 = -3.26 \text{ billion} \]

The equation performed poorly since the prediction represents an extrapolation through time. (The trend might not have been the same before 2003!)
(f) $s = 0.563323$  Yes, same answer except for rounding.

(g) $s_y = 3.67$  Yes, the reduction in standard deviation from 3.67 to 0.5633 is huge!

(h) $R^2 = 98.2\%$

C. T-Bills and Inflation

- 10.6

The scatterplot shows a fairly strong positive linear relationship.

$$\text{T-Bill Return} = 2.964 + 0.6269 \times \text{Inflation}$$
• 10.7
Use Four Steps:
1.
\[ H_A: \beta_1 > 0 \]
\[ H_0: \beta_1 \leq 0 \]

2.
\[ t = \frac{b_1}{SE_{b_1}} = \frac{0.6269}{0.0889} = 7.0517 \]
But “Degrees of freedom” \( df = (n - 2) = 51 - 2 = 49 \implies\) Use \( df = 40 \)
\( \implies P\text{-value} < 0.0005 \)

3. Reject \( H_0 \) since \( P\text{-value} < 0.0005 < 0.05 = \alpha \)

4. There’s enough evidence to show that the rate of return on T-bills is positively related to the rate of inflation.

• 10.8
\[ b_1 \pm t_{n-2}^* \cdot SE(b_1) = b_1 \pm t_{49}^* \cdot SE(b_1) \]
\[ \approx b_1 \pm t_{40}^* \cdot SE(b_1) \]
\[ = 0.6269 \pm (2.021)(0.0889) \]
\[ = 0.6269 \pm 0.1797 = (0.4472, 0.8066) \]
Interpret:
We are 95% confident that a 1% increase in the inflation rate is associated with between 0.4272% and 0.8066% increase in the rate of return on T-bills, on average.

• 10.29
(a) \( \beta_0 \) is the mean rate of return on T-bills in the absence of inflation (with 0% inflation.)
We expect \( \beta_0 > 0 \) simply because no one will want to purchase T-bills from the federal government unless they are paid interest. Someone can earn a 0% return on his/her money simply by putting the money in a safe at home!

(b) From MINITAB,
\[ \hat{\beta}_0 = b_0 = 2.964, \quad SE_{b_0} = 0.445 \]

(c) Four Steps:
1.
\[ H_A: \beta_0 > 0 \]
\[ H_0: \beta_0 \leq 0 \]
2. \[ t = \frac{b_0}{SE_{b_0}} = \frac{2.964}{0.445} = 6.66 \]
\[ \implies P\text{-value} < 0.0005 \]

3. Reject \( H_0 \) since \( P\text{-value} < 0.0005 < 0.05 = \alpha \)

4. There’s enough evidence to show that the mean rate of return on T-bills exceeds 0% in the absence of inflation.

(d) \[ b_0 \pm t^*_{n-2} \cdot SE(b_0) = (2.065, 3.863) \]

Interpret:
We are 95% confident that the mean rate of return on T-bills in the absence of inflation is between 2.07% and 3.86%.

- 10.42

(a) Rounding in MINITAB 17 output differs slightly from MINITAB 16 shown in the textbook but otherwise the answers match.
  
  - From MINITAB “Fit”: \( \hat{y} = 5.28382\% \)
  
  - Does plugging \( x = 3.7 \) into the regression equation verify MINITAB’s answer for \( \hat{y} \)?

(b) \( (1.47629\%, 9.09134\%) \)

- 10.52

(a)

\[ \hat{y} \pm t^*_{n-2,90\%} \cdot SE_{\hat{y}} \]

From \( t \) Table: \( t^*_{n-2,90\%} = t^*_{49,90\%} \approx t^*_{40,90\%} = 1.684 \)

\[ \implies 5.28382 \pm (1.684)(0.264467) = 5.28382 \pm 0.44536 = (4.83846\%, 5.72918\%) \]

(b) Unfortunately, MINITAB contains a programming flaw: It doesn’t provide the “standard error for prediction” \( SE_{\hat{y}} \) for the prediction interval as part of the regression output so there’s no easy way to calculate the answer.

The text does show the formula for \( SE_{\hat{y}} \) on page 559 but we won’t calculate \( SE_{\hat{y}} \) by hand in Stats for Strategy. (Our time is better spent interpreting output.)
D. Earnings for Female Bank Employees

(a) \( x = 94 \) months, \( y = \$389/\text{week} \)

\[ \hat{y} = 349.4 + 0.590x \]
\[ \hat{y} = 349.40 + 0.590(94) \]
\[ \hat{y} = 349.40 + 55.46 = \$404.86 \]

Prediction error = residual = actual − predicted = \( y - \hat{y} = \$389 - \$404.86 = -\$15.86 \)

(b) 
\[ x = 1 \Rightarrow \$349.99 \]
\[ x = 100 \Rightarrow \$408.40 \]
\[ x = 200 \Rightarrow \$467.40 \]
\[ x = 400 \Rightarrow \$585.40 \]

The smallest LOS in the data is seven months and the largest is 228. Therefore predictions for 100 months and 200 months LOS are appropriate since they fall within the range of the \( x \) (predictor) data.

But predictions for one month and 400 months LOS represent extrapolations from the data and so are considered risky forecasts!

(c) 90% confidence interval for \( \beta_0 \):

\[ b_0 \pm t_{n-2} \cdot SE(b_0) = b_0 \pm t_{57} \cdot SE(b_0) \approx b_0 \pm t_{50} \cdot SE(b_0) = 349.4 \pm 1.676(18.1) = 349.4 \pm 30.34 \]

\[ = (319.06, 379.74) \]

Interpret:
We are 90% confident that the average starting salary for female employees who have customer service jobs in Indiana banks is between \$319.06 and \$379.74 per week.

(d) 

\[ \hat{y} \pm t^* \cdot SE_{\hat{y}} = 423.185 \pm (2.009)(15.5530) = (\$391.94, \$454.43) \]

\[ \text{Minitab's answer is} (\$392.040, \$454.329). \text{The hand-calculated answer and Minitab's} \]
\[ \text{answer differ slightly because of different rounding and since Minitab calculates using (the exact) 57 df} \]
\[ \text{instead of 50 df (from the t table.)} \]

(e) \( (\$397.12, \$449.25) \)

(f) 
1. The green (inside) line shows the CI, the purple (outside) line shows the PI.
2. Answers will vary, depending on the "eyeball" estimate.
   The answer is about \( (\$390, \$420) \)
3. The answer is about \( (\$275, \$650) \)
E. Stocks and Bonds

- 10.34
  (a) The least-squares line is
  \[ \hat{y} = 55.58 - (0.1768) x \]
  where
  \[ x = \text{Net cash flow into stock funds, in billions of $} \]
  \[ y = \text{Net cash flow into bond funds, in billions of $} \]
  The fitted line plot suggests that there may be a negative relationship between net inflow into bond funds and net inflow into stock funds.

(b) Four Steps:
  1. Regression model:
     \[ y = \beta_0 + \beta_1 x + \varepsilon \]
     Hypotheses:
     \[ H_A: \beta_1 \neq 0 \]
     \[ H_0: \beta_1 = 0 \]
  2. \[ t = -1.66 \]
     \[ P\text{-value} = 0.111 \] (from MINITAB)
  3. Fail to Reject \( H_0 \) since
     \[ P\text{-value} = 0.111 > 0.05 = \alpha \]
  4. There is not sufficient evidence to show that mean net cash flow into bond funds is linearly related to net cash flow into stock funds.

(c) The plot shows that the relationship between \( x \) and \( y \) is fairly weak (correlation \( r = -0.334 \)). There is some evidence of a negative relationship, but not convincing evidence.

(d) Stop! Significance test has failed! Reliable answers aren’t available.

Note: MINITAB provides the prediction interval \((-77.5701, 153.360)\) but this answer isn’t reliable since the regression significance test has failed.

Sometimes this is described by saying “the predictor variable isn’t statistically significant.” (Please also review Notebook Example 6, page 75.)

(e) Stop! Significance test has failed! Reliable answers aren’t available.
F. Computer Memory

• 10.39
  
  (a)

Questions to answer for (a):

1. \( \hat{\text{DRAM}} = -27,000,000,000 + 13,645,402 \times \text{Year} \)

2. The residuals plot shows a clear pattern. The regression assumptions appear to be violated.

(b)
log(DRAM) = −872.93 + 0.44639 × Year

(c) 90% confidence interval for $\beta_1$:

$$b_1 \pm t^* \cdot SE(b_1) = 0.44639 \pm (2.132)(0.00686)$$

$$= 0.44639 \pm 0.01463$$

$$= (0.43176, 0.46102)$$

Interpret:
We are 90% confident that log(DRAM) increases by between 0.43176 and 0.46102 every year, on average.

(d) A 90% prediction interval from MINITAB for log(DRAM) in 2002 is

$$(20.2971, 21.1921)$$

so a 90% prediction interval for DRAM is

$$e^{20.2971} \text{ to } e^{21.1921} = (653,008,040 \text{ to } 1,598,129,948) \text{ bits}$$
10.38

(a) The fitted regression line is

\[ \hat{y} = -0.01270 + (0.01796) x \]

where

\[ x = \text{Number of beers consumed in 30 minutes} \]
\[ y = \text{Blood alcohol content, in percent} \]

\[ R^2 = 80.0\%. \] The plot shows a strong positive relationship.

(b) Hypothesis Test:

1. \[ H_A: \beta_1 > 0 \]
\[ H_0: \beta_1 \leq 0 \]

2. \[ t = \boxed{7.48} \]

The P-value for this test is half the P-value given by MINITAB:

\[ P\text{-value} = \left(\frac{1}{2}\right) \times 0 = 0 \]

3. Reject \( H_0 \) since \( P\text{-value} = 0 < 0.05 = \alpha \).

4. There is sufficient evidence to show that mean BAC is positively related to the number of beers consumed in 30 minutes.

(c) 90\% confidence interval for \( \beta_1 \):

\[ b_1 \pm t_{n-2}^{*} \cdot SE(b_1) = b_1 \pm t_{14}^{*} \cdot SE(b_1) \]
\[ = 0.01796 \pm (1.761)(0.00240) \]
\[ = 0.01796 \pm 0.00422264 \]
\[ = [0.0137, 0.0222] \]

Interpretation:

We are 90\% confident that each additional beer implies an increase of between 0.0137\% and 0.0222\% in mean BAC.
• 10.55
We predict with 90% certainty that Steve’s BAC is between 0.0400% and 0.1142%. This interval includes values larger than 0.08% so Steve cannot be confident that he won’t be arrested if he drives and is stopped.

H. Predicting Water Quality

• 10.19

(a)

\[ \hat{IBI} = 52.92 + 0.4602 \times \text{Area} \]

• 10.20

(a)

\[ \hat{IBI} = 59.907 + 0.15313 \times \text{Forest} \]
(b) Either predictor variable can be used for regression since both are significant using \( \alpha = 0.10 \) (\( P \)-value for Area = 0.001, \( P \)-value for Forest = 0.061.)

As a tie-breaker, choose Area as the preferred predictor on the basis of either of two equivalent comparisons:

1. Regression standard deviation \( s \) for Area = 16.53 < 17.79 = \( s \) for Forest.
2. \( R^2 \) for Area = 19.9\% > 7.3\% = \( R^2 \) for Forest.

(c) Based on the model using Area, a 90\% prediction interval for IBI is \((47.48, 104.38)\)

(d)

![Fitted Line Plot (Corrected)](image)

\[ \hat{\text{IBI}} = 61.35 + 0.1511 \times \text{Area} \]

An outlier is a data point which is far away from most of the other data points in the scatterplot.

(e)

- After the data correction, Area is no longer a significant predictor at the 10\% level (\( P \)-value= 0.209) and so is disqualified from consideration. However, since Forest is a significant predictor (\( P \)-value = 0.061), it can be used for regression.

- A 90\% prediction interval for IBI using Forest is \((34.33, 94.68)\)