Chapter 3 Review

Part 1: Multiple Choice: Circle the BEST answer for each question.

1. Other things being equal, larger automobile engines consume more fuel. You are planning an experiment to study the effect of engine size (in liters) on the gas mileage (in miles per gallon) of sport utility vehicles. In this study,
   (a) gas mileage is a response variable, and you expect to find a negative association.
   (b) gas mileage is a response variable, and you expect to find a positive association.
   (c) gas mileage is an explanatory variable, and you expect to find a strong negative association.
   (d) gas mileage is an explanatory variable, and you expect to find a strong positive association.
   (e) gas mileage is an explanatory variable, and you expect to find very little association.

2. In a statistics course, a linear regression equation was computed to predict the final-exam score from the score on the first test. The equation was \( y = 10 + 0.9x \) where \( y \) is the final-exam score and \( x \) is the score on the first test. Carla scored 95 on the first test. What is the predicted value of her score on the final exam?
   (a) 85.5  (b) 90  (c) 95  (d) 95.5  (e) none of these

3. In the course described in #2, Bill scored a 90 on the first test and his residual was –2. What was his score on the final exam?
   (a) 89  (b) 91  (c) 93  (d) 90  (e) not enough information

4. The correlation between the heights of fathers and the heights of their (fully grown) sons is \( r = 0.52 \). This value was based on both variables being measured in inches. If fathers' heights were measured in feet (one foot equals 12 inches), and sons' heights were measured in centimeters (2.54 cm equals 1 inch), the correlation between heights of fathers and heights of sons would be
   (a) much smaller that 0.52
   (b) slightly smaller than 0.52
   (c) unchanged: equal to 0.52
   (d) slightly larger than 0.52
   (e) much larger that 0.52

5. It's easy to measure the circumference of a tree's trunk, but not so easy to measure its height. Foresters developed a model for ponderosa pines that they use to predict the tree's height (in feet) from the circumference of its trunk (in inches): \( \ln h = -1.2 + 1.4 \ln C \). A lumberjack finds a tree with a circumference of 60”; how tall does this model estimate the tree to be?
   (a) 5'   (b) 11'     (c) 19'       (d) 83'       (e) 93'

6. In a study of the relationship between SAT math scores (\( x \)) and SAT verbal scores (\( y \)), it was determined that the correlation coefficient was .68. Also, the mean SAT math score was 540 with a standard deviation of 125. The mean SAT verbal score was 498 with a standard deviation of 100. Which of the following is the least squares regression line for this data?

   (a) \( \hat{y} = 204.24 + 0.544x \)
   (b) \( \hat{y} = 66 + 0.8x \)
   (c) \( \hat{y} = -177 + 1.25x \)
   (d) \( \hat{y} = 39 + 0.85x \)
   (e) There is not enough information to determine the equation
One concern about the depletion of the ozone layer is that the increase in ultraviolet (UV) light will decrease crop yields. An experiment was conducted in a greenhouse where soybean plants were exposed to varying levels of UV, measured in Dobson units. At the end of the experiment the yield (kg) was measured. A regression analysis was performed with the results shown in this output:

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std Error</th>
<th>t Ratio</th>
<th>Prob &gt;</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.9800118</td>
<td>0.053774</td>
<td>74.01</td>
<td>&lt;0.001</td>
<td>3.8638348</td>
<td>4.0961838</td>
</tr>
<tr>
<td>UV</td>
<td>-0.046285</td>
<td>0.010741</td>
<td>* hidden</td>
<td>0.008</td>
<td>* hidden</td>
<td>* hidden</td>
</tr>
</tbody>
</table>

7. The least-squares regression line is the line that
(a) minimizes the sum of the squared differences between the actual UV values and the predicted UV values.
(b) minimizes the sum of the squared residuals between the actual yield and the predicted yield.
(c) minimizes the sum of the squared differences between the actual yield and the predicted UV.
(d) minimizes the sum of the squared residuals between the actual UV reading and the predicted UV reading.
(e) minimizes the total variation in the data.

8. Which of the following is correct?
(a) The estimated yield is 3.98 kg when the UV reading is 0 Dobson units.
(b) The predicted yield is 4.3 kg when the UV reading is 20 Dobson units.
(c) If the UV reading increases by 1 Dobson unit, the yield is expected to increase by 0.0463 kg.
(d) If the yield increases by 1 kg, the UV reading is expected to decline by 0.0463 Dobson units.
(e) None of these

9. Using data from a sample of 15 women, this scatterplot shows the relationship between body density (as compared to water density) and skinfold thickness (in mm). Skinfold thickness is used to predict body density. Choose the correct regression equation and correlation for these data.

(a) $density = 1.08 - 0.0003 \cdot (skinfold); \quad r = 0.90$
(b) $density = 1.08 - 0.0003 \cdot (skinfold); \quad r = -0.90$
(c) $density = 1.08 + 0.0003 \cdot (skinfold); \quad r = -0.90$
(d) $density = 1.08 + 0.0003 \cdot (skinfold); \quad r = 0.90$
(e) none of the above


10. This scatterplot shows the overall percentage of on-time arrivals versus overall mishandled baggage per 1000 passengers for the year 2002.

a. Describe the relationship between the on-time percentage and the rate of mishandled baggage.

The relationship between on-time percentage and mishandled baggage rate is positive and linear, with moderate strength. The positive relationship suggests that an airline that tends to be “good” at arriving on time is “bad” at handling baggage, and vice-versa.
b. Which airline has the worst record for mishandled baggage? For being on time?
Northwest has the worst record for mishandled baggage. Alaska has the lowest on-time arrival percentage.

c. United has the highest percentage of on-time arrivals. Estimate that percentage and United’s mishandled baggage rate.
United’s flights arrived on-time about 82% of the time. The airline mishandled about 3.8 bags per 1000 passengers.

11. Here are fat and calorie content for one serving of seven different cereals:

<table>
<thead>
<tr>
<th>Fat (g)</th>
<th>5</th>
<th>2</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calories</td>
<td>120</td>
<td>110</td>
<td>130</td>
<td>90</td>
<td>100</td>
<td>140</td>
<td>150</td>
</tr>
</tbody>
</table>

a. Plot the points so that calories can be predicted from fat.
b. Compute the equation of the least squares regression line.
   Draw the line on your plot.
   \[
   \text{calories} = 110 + 5(fat \text{ grams})
   \]
c. Interpret the slope and \( y \)-intercept in the context of this situation.
   The \( y \)-intercept indicates that one serving of a cereal with 0 fat grams is expected to contain about 110 calories.
   The slope indicates that for every additional gram of fat in a serving of cereal, the calorie count increases by approximately 5.
d. Using your line from part b, estimate the number of calories in one serving of a cereal that contains 4 grams of fat.
   \[
   \text{calories} = 110 + 5(4) = 130 \text{ calories}
   \]
e. Calculate the residual for the cereal that contains 5 grams of fat per serving.
   \[
   \text{residual} = y - \hat{y} = 120 - 135 = -15 \text{ calories}
   \]
f. Verify that the least squares regression line goes through the point of averages, \((\bar{x}, \bar{y})\)
   \[
   \bar{x} = 2, \bar{y} = 120 \rightarrow \text{calories} = 110 + 5(fat \text{ grams}) \rightarrow 120 = 110 + 5(2) \rightarrow 120 = 110 + 10
   \]
   Which is true, so the LSRL does pass through the point \((2, 120)\)
g. What is the sum of the residuals?
   \[
   0
   \]
   \[
   \sum \text{residuals} = -15 - 10 + 10 - 25 - 10 + 25 + 25 = 0
   \]
   (Not that the work shown here is not required as this is a property of the residuals.)
h. Interpret \( r^2 \) in the context of the problem.
   \[
   r^2 = 0.143 \quad \text{The value of} \ r^2 \ \text{indicates that} \ 14.3\% \ \text{of the variation in calorie content can be explained by its linear relationship with fat in a given cereal.}
   \]
i. Interpret \( s \) in the context of the problem if \( s = 21.9089 \).
   When using the LSRL to predict calories from fat content, we will be off by 21.9 calories, on average.
12. How much do consumers in the United States pay for gas? These data show typical prices (in thousands of dollars) and highway gas mileages for a sample of car models commonly sold in the United States. The scatterplot of these data with a regression line is also displayed, followed by a residual plot.

<table>
<thead>
<tr>
<th>Model</th>
<th>Price (thousands $)</th>
<th>MPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chevrolet Suburban LS</td>
<td>33.2</td>
<td>18</td>
</tr>
<tr>
<td>Dodge Caravan EX</td>
<td>24.9</td>
<td>24</td>
</tr>
<tr>
<td>Ford Explorer XLS</td>
<td>24.8</td>
<td>21</td>
</tr>
<tr>
<td>Ford Focus SE</td>
<td>14.5</td>
<td>36</td>
</tr>
<tr>
<td>Ford Taurus</td>
<td>19.2</td>
<td>28</td>
</tr>
<tr>
<td>Honda Civic EX</td>
<td>17.1</td>
<td>38</td>
</tr>
<tr>
<td>Hyundai Elantra GL</td>
<td>11.1</td>
<td>33</td>
</tr>
<tr>
<td>Lexus GS300</td>
<td>36.1</td>
<td>25</td>
</tr>
<tr>
<td>Saturn L200</td>
<td>17.8</td>
<td>33</td>
</tr>
<tr>
<td>Toyota Corolla CE</td>
<td>14.4</td>
<td>41</td>
</tr>
<tr>
<td>Volvo S60</td>
<td>29.7</td>
<td>28</td>
</tr>
</tbody>
</table>

a. Interpret the slope of the regression line in the context of this problem. The slope is -0.689. It indicates for every increase of $1000 in the price of a car, that car’s fuel economy goes down by about .689 miles per gallon (highway).

b. Identify the correlation coefficient for the relationship shown. Explain the meaning of that number.
   \[ r = \sqrt{r^2} = \sqrt{0.61} = 0.781, \text{ but because the relationship is negative, the correlation should be as well.} \]
   \[ r = -0.781 \]
   A correlation of -0.781 indicates that the linear relationship between price and mileage is strong (or moderately strong) and negative.

c. Is fitting a regression line appropriate for the pattern displayed in the scatterplot? Explain by inspecting the original plot as well as the residual plot.
   Yes, a regression line is appropriate. The original plot shows a negative linear trend, and the residual plot shows a random scattering of points and no curved patterns like smiles or frowns, which might indicate that a nonlinear model would be more appropriate.

d. Does the data indicate that raising the price of a car will cause it to have lower fuel economy?
   No. Correlation does not imply causation.

e. What lurking variable might account for the negative association between price and gas mileage?
   “Size of the car” might contribute to the negative association. Larger cars tend to cost more than smaller cars, but they also tend to be heavier, thus requiring more fuel to move them.

f. Which auto produces the largest positive residual? Compute that residual. (Do not round your answer.)
   On the TI-Nspire, \( \text{residual} = y - \hat{y} = \text{mpg} - f1(\text{price}) \) (or whatever names you used for the MPG and price lists)
   The Toyota Corolla CE’s residual is \( y - \hat{y} = (41 - 34.833) = 6.166803954969 \)
   (6.16680395 or possibly even 6.1668 depending on how many digits you read from your calculator)
13. A friend is planning to sell her used Honda so she gathers data on prices (in thousands of dollars) and year (let 79 represent 1979) for eleven cars similar to hers. A scatterplot of the data is shown. The least squares regression line through these data has the equation

\[ \hat{\text{Price}} = -59.5 + 0.735 \cdot \text{Year} \]

a. Interpret the slope of the regression line in the context of this problem.

The price for used Hondas tends to rise, on average, about $735 per model year.

b. The data point for the 1979 Honda is removed from the data set, and a new regression model is calculated. State whether the slope increases or decreases.

The slope increases.

c. Your friend’s Honda is a 1988. Use the regression model to establish a selling price for her car.

The equation becomes \[ \hat{\text{Price}} = -59.5 + 0.735(88) \], or 5.18. Therefore, the estimated price of a used 1988 Honda would be $5180.

d. Looking at these data, do you think the selling price established in part c is too high, too low, or about right? Explain.

The regression line is above the points near the year 1988, which suggests that the estimate might be a bit higher than it should be.

Conversely, both the original regression line and the modified regression line (without the 1979 Honda) seem to predict about the same value for a 1988 Honda. So that estimate would hold approximately the same with either line, making it about right.

(\text{NOTE: the first argument is quite a bit stronger than the second one, but you should only make one argument. Pick a position and defend it})

14. One weekend, a statistician notices that some of the cars in his neighborhood are very clean and others are quite dirty. He decides to explore this phenomenon, and asks 15 of his neighbors how many times they wash their cars each year and how much they paid in car repair costs last year. His results are in the table below:

<table>
<thead>
<tr>
<th>x = number of car washes per year</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>y = repairs costs for last year</td>
<td>$955.30</td>
<td>$323.50</td>
</tr>
</tbody>
</table>

The correlation for these two variables is \( r = -0.71 \), and a scatterplot reveals a roughly linear relationship.

a. Find the equation of the least-squares regression line (with \( y \) as the response variable).

\[
\text{Slope} = b = r \left( \frac{s_y}{s_x} \right) = -0.71 \left( \frac{323.5}{3.78} \right) = -60.763; \\
\text{Y-intercept} = a = \bar{y} - bx = 955.3 - (-60.763)(6.4) - 1344.18. \text{ So } \hat{y} = 1344.18 - 60.763x.
\]

where \( \hat{y} \) is the predicted repair costs and \( x \) is the number of car washes per year.

b. Interpret the coefficient of determination.

\( r^2 = 0.5041 \)

50.41\% of the variation in repair costs last year can be explained by its linear relationship with number of car washes each year.

c. Based on these data, can we conclude that washing your car frequently will reduce repair costs? Explain.

No. Since this was not a controlled experiment, there could be lurking variables that are responsible for the association observed here. Perhaps the frequency with which drivers wash their cars is confounded with other good car-maintenance habits, such as changing the car’s oil frequently. Correlation does not imply causation.
For a laboratory experiment, scientists induced a skin infection in a rat and recorded the growth of the discolored skin patch every three days over a six-week period. Ten of the recordings are shown here.

<table>
<thead>
<tr>
<th>Days After Induction</th>
<th>14</th>
<th>17</th>
<th>20</th>
<th>23</th>
<th>26</th>
<th>29</th>
<th>32</th>
<th>35</th>
<th>38</th>
<th>41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter of Patch (mm)</td>
<td>1.9</td>
<td>3.1</td>
<td>4.6</td>
<td>7.4</td>
<td>11.3</td>
<td>16.4</td>
<td>25.4</td>
<td>38.0</td>
<td>55.9</td>
<td>81.9</td>
</tr>
</tbody>
</table>

a. Plot diameter versus days.

b. What transformation should Gillian use on these data before she fits a linear equation? What does this tell you about the original data?

She should use a Log Transformation
Plot: \((\text{days, ln(diameter)})\)

This indicates the data might follow an exponential model.

c. Find the equation of the regression line for \(\ln(\text{diameter})\) vs. days.

\[
\ln(\text{diam}) = -1.237 + 0.1389(\text{days})
\]

The graph of the transformed data is also shown at right.

Based on the plot of the residuals (shown at right, below), there is still a pattern of curvature, indicating that this model is not appropriate either. Perhaps another transformation could do better. (or) Perhaps another model would be more appropriate.

d. Use your model from part c to estimate the size of the patch after 33 days, and after 55 days. Do you have the same level of confidence in both estimates? Explain.

\[
\ln(\text{diam}) = -1.237 + 0.1389(33) \\
\ln(\text{diam}) = 3.346 \\
\hat{\text{diam}} = 28.397
\]

After 33 days, the patch is expected to measure about 28.4 mm.

\[
\ln(\text{diam}) = -1.237 + 0.1389(55) \\
\ln(\text{diam}) = 6.40175 \\
\hat{\text{diam}} = 602.9
\]

After 55 days, the patch is expected to measure about 602.9 mm in diameter.
We should have much more confidence in the 33-day estimate, as it is interpolation with a very strong ($r = 0.99954$) model, while the 55-day estimate is extrapolation, far outside our existing data range.

e. From examining the residuals, does it appear that your estimate for 55 days will be too high, too low, or about right?

Based on the curvature seen in the residual plot, we would expect the estimate for 55 days to be too high as the actual observation will likely lie below the LSRL.