1. Inches of rainfall during the growing season is related to the yield of corn in bushels per acre. Which of these variables should be considered the explanatory variable and which should be considered the response variable?

Explanatory: Rainfall  
Response: Yield

2. Given an example of two variables that have a strong, positive, approximately linear relationship but for which it seems certain that neither variable “causes” the other. Lots of examples are possible. But a correct answer lists two variables that meet all four conditions. Several people gave examples that might reasonably be said to be causally related. Some people gave examples of variables but no indication of what individuals these variables are measured on so that it was not possible to see that the relationship was strong, positive, and/or linear.

3. There are 83 counties in the state of Michigan. Suppose that we wished to select a sample of 10,000 Michigan residents. One method would be to take a stratified random sample using the counties as strata. How would we do this? (You may suppose that we knew the number of residents in each county.)

Select a simple random sample within each county. The size of these simple random samples could be proportional to the population of the counties or some fixed size (although in the latter case one has to be careful how to amalgate the data(.

4. Consider \( n \) pairs of numbers \((x_1, y_1), \ldots, (x_n, y_n)\). Suppose that the mean and standard deviation of the \( x \) values are \( \bar{x} = 5 \) and \( s_x = 4 \) respectively and that the mean and standard deviation of the \( y \) values are \( \bar{y} = 10 \) and \( s_y = 10 \) respectively. Which of the following could be the least squares regression line?

(a) \( \hat{y} = -5 + 3x \)  
(b) \( \hat{y} = 3x \)  
(c) \( \hat{y} = 5 + 2.5x \)  
(d) \( \hat{y} = 8.5 + 0.3x \)  
(e) \( \hat{y} = 10 + 0.4x \)

The regression line must go through the point \((\bar{x}, \bar{y})\) and can have slope no larger than \( \frac{s_y}{s_x} \).

5. Suppose that I want to know what Calvin College students think about a controversial issue.

(a) What is the purpose of choosing a simple random sample rather than simply asking the first 100 students that I see in Johnny’s?

An SRS has a good chance to be representative of the whole population. Sampling from Johnny’s might lead to bias of various kinds (towards commuter students?).

(b) In this situation, the collection of all Calvin students is called the population.
6. In the computation of a least squares regression line, what does the phrase “least squares” refer to? (i.e., which squares are least?)
The least squares line is the one that minimizes the sums of the squares of the residuals.

7. SAT scores on the mathematics test (SATM) and verbal test (SATV) are in the range 200-800. The SATM and SATV scores of all Calvin seniors who have such scores is in the scatter plot below.

![Scatter plot of SATM vs. SATV scores](image)

Based on the scatter plot, the most likely value for the correlation between SATM and SATV is (choose one)

(a) -.9  (b) -.5  (c) 0  (d) .5  (e) .9

\( r = .5 \) is a moderately positive linear relationship.

8. A regression is performed using the data in the preceding problem. The resulting equation is

\[ SATV = 264 + 0.55 \text{ SATM} \]

If a student scores 500 on the SATM, what is the predicted SATV score? \[264 + 0.55(500) = 539\].

9. The population, number of physicians, number of televisions, and average life expectancy of 41 countries is summarized in the following two scatter plots. In each graph below, the response variable is life expectancy. In the first, the explanatory variable is the number of people per television. In the second, the explanatory variable is the number of people per physician.

(a) Compare the two relationships as to their form, strength, and direction.

Various descriptions of the two relationships are possible. But at minimum the answer must be a comparison. That is two lists of characteristics is not enough. Compare means to use words like stronger, different, the same, etc.
Most people noted that neither plot is linear and both have the same curvilinear relationship and it is negative. People differed as to which is stronger with defensible reasons for each answer. People who thought the relationship in each case was linear could defend this by saying there are a lot of outliers.

(b) The first graph seems to suggest that we could increase the average life expectancy in certain countries by exporting televisions to those countries. Do you agree? Use the concepts introduced in the book to explain your answer. Obviously this seems silly. Televisions don’t directly affect life expectancy. (Although they might contribute to decreasing it due to couch potato syndrome.) What’s going on here is that both variables are affected by the relative economic wealth of the given countries.

(c) Suppose that we compute a least squares regression line for the data in the second of the plots above. Plot on the axes below a rough sketch of a plot of the residuals.

The answer to this question depends on where you think the least squares line is. Credit was given for any plot consistent with some even somewhat plausible regression line. But not all residuals can be positive and not all residuals can be negative! The raw data is available on the website and you can see for yourself what the plot should look like.

10. The fifty states in the United States are listed here:

<table>
<thead>
<tr>
<th>Alabama</th>
<th>Alaska</th>
<th>Arizona</th>
<th>Arkansas</th>
<th>California</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorado</td>
<td>Connecticut</td>
<td>Delaware</td>
<td>Florida</td>
<td>Georgia</td>
</tr>
<tr>
<td>Hawaii</td>
<td>Idaho</td>
<td>Illinois</td>
<td>Indiana</td>
<td>Iowa</td>
</tr>
<tr>
<td>Kansas</td>
<td>Kentucky</td>
<td>Louisiana</td>
<td>Maine</td>
<td>Maryland</td>
</tr>
<tr>
<td>Massachussets</td>
<td>Michigan</td>
<td>Minnesota</td>
<td>Mississippi</td>
<td>Missouri</td>
</tr>
<tr>
<td>Montana</td>
<td>Nebraska</td>
<td>Nevada</td>
<td>New Hampshire</td>
<td>New Jersey</td>
</tr>
<tr>
<td>New Mexico</td>
<td>New York</td>
<td>North Carolina</td>
<td>North Dakota</td>
<td>Ohio</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>Oregon</td>
<td>Pennsylvania</td>
<td>Rhode Island</td>
<td>South Carolina</td>
</tr>
<tr>
<td>South Dakota</td>
<td>Tennessee</td>
<td>Texas</td>
<td>Utah</td>
<td>Vermont</td>
</tr>
<tr>
<td>Virginia</td>
<td>Washington</td>
<td>West Virginia</td>
<td>Wisconsin</td>
<td>Wyoming</td>
</tr>
</tbody>
</table>

(a) Describe carefully how you would use Table B to choose a simple random sample of five states from these 50. Describe the method carefully enough so that I could choose a random sample using that method.

Number the states using two-digit numbers. (Say from 01 to 50). Choose a line in Table B to start at. Choose successive two-digit numbers from the list until five different numbers in the range 01-50 are chosen. (We will need to discard all numbers from 51-99 and 00 and all duplicates.) Choose the states with those numbers.

If your answer in this part didn’t give me enough instruction to verify that you had correctly implemented the strategy in part (b), it was deficient.
(b) Choose a simple random sample of five states using the method you described in part (a).
Many answers are possible. The grader went cross-eyed reading various lines of Table B.

11. For each state, we know the percentage of the residents that are born in the US (usborn) and the average household income of residents of the state (income). A regression is performed and we obtain the following

Simple linear regression results:
Dependent Variable: income
Independent Variable: usborn
income = 68642.24 - 46018.64 usborn
Sample size: 51
R (correlation coefficient) = -0.5781
R-sq = 0.33419776

and the data is plotted with the regression line below.

(a) On the graph, circle the observation that has the largest residual.
The highest one.

(b) On the graph, put a square around the observation that is the most influential on the regression analysis.
The one furthest left.
(c) Note that the variable \texttt{usborn} is a number between 0 (0\%) and 1 (100\%). Complete the following: For each increase in the percentage of usborn citizens by 0.1 (10\%), the average household income \boxed{\text{decreases}} (answer increases or decreases) by $4,601$. (answer an amount).

(d) The number labeled \texttt{R-sq} is usually read as a percentage; i.e., 33\%. Write a sentence using 33\% that uses this number to make a conclusion about the relationship between the two variables.

33\% of the variation in income from state to state is explained by the variation in the percentage of usborn citizens in the state.