Chapter 4

Scatterplots and Correlation
Explanatory and Response Variables

- Interested in studying the relationship between two variables by measuring both variables on the same individuals.
  - a *response variable* measures an outcome of a study
  - an *explanatory variable* explains or influences changes in a response variable
  - sometimes there is no distinction
Question

In a study to determine whether surgery or chemotherapy results in higher survival rates for a certain type of cancer, whether or not the patient survived is one variable, and whether they received surgery or chemotherapy is the other. Which is the explanatory variable and which is the response variable?
Scatterplot

- Graphs the relationship between two quantitative (numerical) variables measured on the same individuals.
- If a distinction exists, plot the explanatory variable on the horizontal (x) axis and plot the response variable on the vertical (y) axis.
Scatterplot

Relationship between mean SAT verbal score and percent of high school grads taking SAT

This is the point for Colorado, where 28% took the SAT and the mean verbal score was 543.
To add a categorical variable, use a different plot color or symbol for each category.

Southern states highlighted
Scatterplot

- Look for *overall pattern* and *deviations* from this pattern
- Describe pattern by *form, direction*, and *strength* of the relationship
- Look for *outliers*
Scatterplot

- **Form**

- [Image of scatterplots showing different forms]

BPS - 5th Ed.  Chapter 4  8
Scatterplot

- Direction
Scatterplot

- **Strength**

![Scatterplot diagrams](image)
Linear Relationship

Some relationships are such that the points of a scatterplot tend to fall along a straight line -- linear relationship.
Direction

◆ Positive association
  – above-average values of one variable tend to accompany above-average values of the other variable, and below-average values tend to occur together

◆ Negative association
  – above-average values of one variable tend to accompany below-average values of the other variable, and vice versa
Examples

From a scatterplot of college students, there is a *positive association* between verbal SAT score and GPA.

For used cars, there is a *negative association* between the age of the car and the selling price.
Examples of Relationships

- Income vs. Health Status Measure
- Age vs. Health Status Measure
- Education Level vs. Age
- Mental Health Score vs. Physical Health Score
Measuring Strength & Direction of a Linear Relationship

- How closely does a non-horizontal straight line fit the points of a scatterplot?
- The correlation coefficient (often referred to as just correlation): \( r \)
  - measure of the strength of the relationship: the stronger the relationship, the larger the magnitude of \( r \).
  - measure of the direction of the relationship: positive \( r \) indicates a positive relationship, negative \( r \) indicates a negative relationship.
Correlation Coefficient

- special values for $r$:
  - a perfect positive linear relationship would have $r = +1$
  - a perfect negative linear relationship would have $r = -1$
  - if there is no linear relationship, or if the scatterplot points are best fit by a horizontal line, then $r = 0$
  - *Note*: $r$ must be between $-1$ and $+1$, inclusive

- both variables must be quantitative; no distinction between response and explanatory variables

- $r$ has no units; does not change when measurement units are changed (ex: ft. or in.)
Examples of Correlations

- Correlation $r = 0$
- Correlation $r = -0.3$
- Correlation $r = 0.5$
- Correlation $r = -0.7$
- Correlation $r = 0.9$
- Correlation $r = -0.99$
Examples of Correlations

- Husband’s versus Wife’s ages
  - $r = 0.94$
- Husband’s versus Wife’s heights
  - $r = 0.36$
- Professional Golfer’s Putting Success: Distance of putt in feet versus percent success
  - $r = -0.94$
Not all Relationships are Linear
Miles per Gallon versus Speed

- Linear relationship?
- Correlation is close to zero.

\[ y = -0.013x + 26.9 \]
\[ r = -0.06 \]
Not all Relationships are Linear
Miles per Gallon versus Speed

- Curved relationship.
- Correlation is misleading.
Problems with Correlations

- Outliers can inflate or deflate correlations (see next slide)
- Groups combined inappropriately may mask relationships (a third variable)
  - groups may have different relationships when separated
For each scatterplot above, how does the outlier affect the correlation?

A: outlier decreases the correlation
B: outlier increases the correlation
Correlation Calculation

- Suppose we have data on variables $X$ and $Y$ for $n$ individuals:
  
  $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$

- Each variable has a mean and std dev:
  
  $(\bar{x}, s_x)$ and $(\bar{y}, s_y)$ (see ch. 2 for $s$)

$$
r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
$$
Case Study

Per Capita Gross Domestic Product and Average Life Expectancy for Countries in Western Europe
## Case Study

<table>
<thead>
<tr>
<th>Country</th>
<th>Per Capita GDP (x)</th>
<th>Life Expectancy (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>21.4</td>
<td>77.48</td>
</tr>
<tr>
<td>Belgium</td>
<td>23.2</td>
<td>77.53</td>
</tr>
<tr>
<td>Finland</td>
<td>20.0</td>
<td>77.32</td>
</tr>
<tr>
<td>France</td>
<td>22.7</td>
<td>78.63</td>
</tr>
<tr>
<td>Germany</td>
<td>20.8</td>
<td>77.17</td>
</tr>
<tr>
<td>Ireland</td>
<td>18.6</td>
<td>76.39</td>
</tr>
<tr>
<td>Italy</td>
<td>21.5</td>
<td>78.51</td>
</tr>
<tr>
<td>Netherlands</td>
<td>22.0</td>
<td>78.15</td>
</tr>
<tr>
<td>Switzerland</td>
<td>23.8</td>
<td>78.99</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>21.2</td>
<td>77.37</td>
</tr>
</tbody>
</table>
## Case Study

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$(x_i - \bar{x})/s_x$</th>
<th>$(y_i - \bar{y})/s_y$</th>
<th>(\frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y})</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.4</td>
<td>77.48</td>
<td>-0.078</td>
<td>-0.345</td>
<td>0.027</td>
</tr>
<tr>
<td>23.2</td>
<td>77.53</td>
<td>1.097</td>
<td>-0.282</td>
<td>-0.309</td>
</tr>
<tr>
<td>20.0</td>
<td>77.32</td>
<td>-0.992</td>
<td>-0.546</td>
<td>0.542</td>
</tr>
<tr>
<td>22.7</td>
<td>78.63</td>
<td>0.770</td>
<td>1.102</td>
<td>0.849</td>
</tr>
<tr>
<td>20.8</td>
<td>77.17</td>
<td>-0.470</td>
<td>-0.735</td>
<td>0.345</td>
</tr>
<tr>
<td>18.6</td>
<td>76.39</td>
<td>-1.906</td>
<td>-1.716</td>
<td>3.271</td>
</tr>
<tr>
<td>21.5</td>
<td>78.51</td>
<td>-0.013</td>
<td>0.951</td>
<td>-0.012</td>
</tr>
<tr>
<td>22.0</td>
<td>78.15</td>
<td>0.313</td>
<td>0.498</td>
<td>0.156</td>
</tr>
<tr>
<td>23.8</td>
<td>78.99</td>
<td>1.489</td>
<td>1.555</td>
<td>2.315</td>
</tr>
<tr>
<td>21.2</td>
<td>77.37</td>
<td>-0.209</td>
<td>-0.483</td>
<td>0.101</td>
</tr>
</tbody>
</table>

$\bar{x} = 21.52$  $\bar{y} = 77.754$

$s_x = 1.532$  $s_y = 0.795$

$\sum = 7.285$
Case Study

\[
r = \frac{1}{n - 1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
\]

\[
= \left( \frac{1}{10 - 1} \right) (7.285)
\]

\[
= 0.809
\]