Chapter 7: Scatterplots, association, and correlation

This chapter introduces methods of analyzing the relationship between two quantitative variables\(^1\). The data of interest consist of pairs of observations; one element of the pair is the value of the first quantitative variable and the second element is the value of the second quantitative variable. The scatterplot is commonly used to visually analyze the relationship between two quantitative variables. The pairs are plotted by assigning one axis to each variable.

Example: Body mass index (BMI) is the ratio of a person’s height to (squared) weight, and according to the Centers for Disease Control website: *BMI provides a reliable indicator of body fatness for most people and is used to screen for weight categories that may lead to health problems.* BMI is defined as

\[
\text{BMI} = \frac{\text{Weight}}{\text{Height}^2}.
\]

However, a heavily muscled individual with relatively little body fat may be incorrectly categorized as overweight using BMI. Further, muscle tissue weights 15% more than fat tissue, and may not be a good measure of whether a person is a risk for certain diseases related to being overweight.\(^2\) *Percent body fat* is a better measure of an individual’s fitness level as it reflects an individual’s body composition without regard to the individual’s height or weight.\(^3\) Percent body fat is relatively expensive and time-consuming to measure, and so BMI is commonly used as a surrogate (or in place of) percent body fat as a measure of body fatness. A question of interest then, is the type and strength of association between BMI and percent body fat.

The Australian athletes data set was collected in a study of how data on various characteristics of the blood varied with sport body size and sex of the athlete. In particular, the authors\(^4\) were interested in determining whether blood hemoglobin concentrations of athletes in endurance-related events differ from those in power-related events. This data set can be used to investigate several interesting relationships and comparisons.

---

\(^1\)Chapter 3 addressed methods of analyzing the relationship between two categorical variables

\(^2\)There is abundant evidence that a number of Cardiovascular risk factors are related to visceral fat, though the strength of the relationships often is unclear.

\(^3\)Dual energy X-ray absorptiometry uses X-rays of two different energies to scan the body, one of which is absorbed more strongly by fat than the other. The difference in absorption reflects the amount of fat relative to other tissues and leads to a computation of overall body composition.

The figure to the right shows the relationship between percent body fat and BMI. Individuals are identified by gender. Is there a clear relationship between percent body fat and BMI? Is there a difference between females and males? How well does BMI predict percent body fat?

Skinfold thickness is a relatively inexpensive but accurate measure of adipose fat tissue (subcutaneous fat). Fat tissue that is not adipose is visceral fat, also known as organ fat or intra-abdominal fat. It’s located inside the abdominal cavity and has been found to be associated with the incidence of cardiovascular diseases.

The figure to the right shows that the relationship between percent body fat and skinfold thickness is strong and linear, though the relationship differs by gender. Given the same percent body fat, women tend to have less subcutaneous fat than men. Presumably, women have somewhat more visceral fat than men when percent body fat is the same.

---

5 Yes, but it’s weak.
6 Yes.
7 Poorly.
Further examples

1. Is there an association between island size and the number of reptile or amphibian species an island can support? The figure plots the number of species against island area for 7 West Indies islands. The figure shows an association and that the relationship between island size and number of species is curvilinear.

2. The figure to the right shows weights of 17 anorexia patients before and after an experiment assessing the efficacy of a behavior modification treatment. These patients were in the control group and received a placebo. A line with slope 1 and intercept 0 is graphed on the figure. Is there an association between pre- and post-experiment weight? 

---

9No.
3. The figure to the right shows IQ and language test scores of 2287 eighth-grade pupils (aged about 11) in 132 classes in 131 schools in the Netherlands. The association is positive and moderate in strength.

Describing association between variables:
There are four principal properties of bivariate relationships.

1. Direction of the association specifies the direction in which one variable changes as the second variable changes.
   (a) Positive association: two variables are positively associated if the values of one variable tend to increase with the values of the second variable. In this case, large values of one variable generally correspond to larger value of the second variable. Examples?^
   (b) Negative association: two variables are negatively associated if the values of one variable tend to increase as the values of the second variable decrease. In other words, large values of one variable tend to be associated with smaller values of the second variable. The figure to the right illustrates. The data shown are total fertility rate (children per woman) and percent of contraceptors among married women of childbearing age for a number of developing countries.^

---


^Percent body fat and skinfold thickness, language test and IQ scores, number of species and island area.

(c) Other types of association: Two variables may be associated in both a positive and negative fashion over different ranges of values of the variables. Philippine TB rates versus time follow a nonlinear relationship.

(d) No association: Two variables are not associated if there is no apparent association (positive, negative, or otherwise) present in the scatterplot. An example is provided by the anorexia experiment data. A plot of pre- and post-experiment weights of the placebo group subjects showed no relationship between pre- and post-experiment weights.

2. *Form*: describes the shape of the association. The shape of association can be visualized if a smooth is fit through the plotted points.

(a) *Linear*: The relationship between two quantitative variables is *linear* points generally follow a straight line. The plot to the left shows record times in 1984 for 35 Scottish hill races. A smooth shows that the form of the association is roughly linear. Another example of linear trend is the relationship between fertility and contraceptive use.

(b) *Nonlinear*: If the relationship is not linear, then it nonlinear. Common forms of nonlinear relationship are *exponential* and *quadratic*. The figure below and left shows an exponential (decay) relationship. The data are concentrations of the chemical GAGurine in the urine of 314 children aged from zero to seventeen years. The aim of the study was to produce a chart to help a pediatrician to assess if a child's GAG concentration is normal.
3. **Strength**: The strength of the relationship is related to the degree of spread of the points away from an apparent trend. The figure to the right and above compares red\(^{13}\) and white blood cell\(^{14}\) concentrations in the blood of the Australian athletes. There is no association between the two variables.

   Less scatter (stronger association) more clearly reveals trends or patterns. If there is too much scatter, an existing relationship between the variables may not be visible.

   The panel plot (right and above) shows (clockwise) strong positive, weak negative, moderate to strong positive, and moderate negative association.

4. **Unusual Features** are departures from linear patterns (e.g., Philippine TB rates versus time) and unusual data pairs not fitting the overall relationship between the variables.

   The figure to the below shows an association between sugar and calorie content of breakfast cereals. There are two cereals (Grape Nuts and Great Grains Pecan\(^{15}\)) that are considered to be outliers since they do not fit the general pattern of association.

---

\(^{13}\)Red blood cells are a vertebrate organism’s principal means of delivering oxygen to the body tissues.

\(^{14}\)White blood cells are cells of the immune system involved in defending the body against infectious disease and foreign materials.

\(^{15}\)Their large calorie contents is attributable to relatively large fat contents.
Plotting: The y-axis (vertically plotted) variable is sometimes viewed as explaining, or responding to the x-axis (horizontally plotted variable). Consider the following questions:

1. Are countries with lower highway speed limits more likely to have lower highway death rates?

2. How well can mean January temperature of a US city be estimated by its latitude?

3. Are yearly carbon dioxide measurements from the atmosphere related to average yearly temperatures?

Each question is implying specific roles for each variable. One variable explains or controls the values of the second variable. When this separation of variables makes sense (and it does not always make sense\textsuperscript{16}), then the following terms are used:

1. **Response**: the variable to be predicted or explained. This variable is placed on the vertical or y-axis. Example: mean January temperature is partly determined by latitude.

2. **Explanatory or predictor variable**: the variable that explains or can be used to predict the values of the response variable. This variable is placed on the horizontal (or x-axis). Example: latitude.

The convention in statistics is to always name the vertically plotted variable first when describing a plot. For example, the plot above shows calories per portion plotted against sugar per portion.

**Correlation**: The correlation coefficient $r$ is a measure of the strength of the linear association between two quantitative variables. The data are observation pairs $(x_1, y_1), \ldots, (x_n, y_n)$, and the correlation coefficient is

$$r = \frac{1}{n - 1} \sum_{i=1}^{n} \left( \frac{x_i - \overline{x}}{s_x} \right) \left( \frac{y_i - \overline{y}}{s_y} \right).$$  \hspace{1cm} (1)

- $r$ sometimes called *Pearson’s correlation* to distinguish it from other measures of association; however, the term *correlation coefficient* in statistics refers specifically to $r$.

\textsuperscript{16}For example, height and weights of children.
Remarks

- Equation (1) firsts calculates the standardized score for \( x_i \) and then for \( y_i \). The product of the standardized \( z \)-scores (which measure the scaled distance of the observation to its mean) are then summed and averaged. Another expression for \( r \) is

\[
 r = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}
\]

where \( z_{x_i} = \frac{x_i - \bar{x}}{s_x} \) and \( z_{y_i} = \frac{y_i - \bar{y}}{s_y} \).

- If two variables are positively associated, then \( x_i \) and \( y_i \) will tend to both be larger their means, or both be less than their means, producing a positive product of \( z \)-scores. Hence, the average of the products will be positive when there is a positive association between the variables.

- The reverse argument holds for negatively associated variables.

\[0.5 1.0 1.5 2.0 2.5 3.0 3.5\]
\[0 1 2 3 4 5\]

\[\begin{array}{c}
0.5 \\
1.0 \\
1.5 \\
2.0 \\
2.5 \\
3.0 \\
3.5
\end{array}\]

\[\begin{array}{c}
0 \\
1 \\
2 \\
3 \\
4 \\
5 \\
6
\end{array}\]

Example: The scatterplot to the right is of mean February temperature against latitude for 12 US cities. The contribution of each city towards \( r \) is tabulated below. For Houston, the contribution is broken down:

\[
z_{x} = \frac{30 - 37.9}{6.36} = -1.245
\]
\[
z_{y} = \frac{54 - 42.0}{17.157} = .699 \Rightarrow z_{x} \cdot z_{y} = -.870.
\]

The contribution of each city towards \( r \) is tabulated below.
<table>
<thead>
<tr>
<th>City</th>
<th>Latitude</th>
<th>z-score</th>
<th>Temp.</th>
<th>z-score</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miami</td>
<td>26</td>
<td>-1.874</td>
<td>69</td>
<td>1.574</td>
<td>-2.949</td>
</tr>
<tr>
<td>Houston</td>
<td>30</td>
<td>-1.245</td>
<td>54</td>
<td>.699</td>
<td>-0.870</td>
</tr>
<tr>
<td>Phoenix</td>
<td>33</td>
<td>-0.773</td>
<td>58</td>
<td>0.933</td>
<td>-0.721</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>34</td>
<td>-0.616</td>
<td>58</td>
<td>0.933</td>
<td>-0.574</td>
</tr>
<tr>
<td>Raleigh</td>
<td>36</td>
<td>-0.301</td>
<td>41</td>
<td>-0.058</td>
<td>0.018</td>
</tr>
<tr>
<td>San Francisco</td>
<td>38</td>
<td>0.013</td>
<td>52</td>
<td>0.583</td>
<td>0.008</td>
</tr>
<tr>
<td>Washington</td>
<td>39</td>
<td>0.170</td>
<td>34</td>
<td>-0.466</td>
<td>-0.079</td>
</tr>
<tr>
<td>New York</td>
<td>41</td>
<td>0.485</td>
<td>34</td>
<td>-0.466</td>
<td>-0.226</td>
</tr>
<tr>
<td>Boston</td>
<td>42</td>
<td>0.642</td>
<td>30</td>
<td>-0.699</td>
<td>-0.449</td>
</tr>
<tr>
<td>Minneapolis</td>
<td>45</td>
<td>1.114</td>
<td>18</td>
<td>-1.399</td>
<td>-1.558</td>
</tr>
<tr>
<td>Eugene</td>
<td>44</td>
<td>0.957</td>
<td>44</td>
<td>0.117</td>
<td>0.112</td>
</tr>
<tr>
<td>Duluth</td>
<td>47</td>
<td>1.428</td>
<td>12</td>
<td>-1.749</td>
<td>-2.497</td>
</tr>
<tr>
<td>Mean</td>
<td>37.917</td>
<td></td>
<td>42.000</td>
<td></td>
<td>-9.788</td>
</tr>
<tr>
<td>Std.Dev.</td>
<td>6.360</td>
<td></td>
<td>17.157</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From the leftmost column,

\[ r = \frac{\sum z_x z_y}{n - 1} = \frac{-9.79}{11} = -.89. \]  \hspace{2cm} (2)

**Properties of correlation**

1. \( r \) measures only the strength of linear association between two variables.

2. The value of \( r \) is always between \(-1\) and 1.
   - If \( r > 0 \), then the association is between variables is positive.
   - If \( r < 0 \), then the association is between variables is negative.
   - If \( r \approx 0 \), then there is no association between variables.

3. If \( r = 1 \) or \( r = -1 \), then there is perfect positive or negative linear association, respectively, between variables.

4. The value of \( r \) is unitless, and so the meaning of \( r \) as a measure of correlation (as discussed above) holds for any pair of quantitative variables.
5. There is no distinction between the \( x \) and \( y \) variable in terms of one variable explaining the other variable.

*StatCrunch:* Correlation coefficients can be computed by utilizing the Stat - Summary Stats - Correlation menus.

*More examples:* Describe the relationship between the two variables in each of these scatterplots:

**Missoula Monthly Average Temperatures**

(degrees Fahrenheit)

**Distance from Sun vs. Planet Position**

(for all 9 planets)

**Remarks**

- The plot on the left and above illustrates that the correlation coefficient is a measure of linear association. \( r = .16 \) does not imply that a relationship does not exist between variables.

- The planet distance plot illustrates that a curved (but monotone) relationship may have a large correlation coefficient provided that the points lie close to a line. For these data, \( r = .91 \).

- In some instances, a nonlinear relationship can be linearized by transforming one or both of the variables. Island area and number of species has been transformed to the natural logarithm scale to produce a linear relationship.

*Effect of outliers on \( r \):* The correlation coefficient resistant to outliers is sensitive to outliers. Consider the two scatterplots below. Guess what the correlations would be with and without the outlier in each plot.

56
There are several measures of association which are resistant to the effects of outliers. Two of the more common such measures are Kendall’s $\tau$ and Spearman’s $\rho$. Both of these measure the degree of monotonic association between two quantitative variables.

**Causation:** Two variables have a cause-and-effect relationship if changes in one variable cause changes in the other variable. **Examples:** The change in the concentration of a catalyst causes changes in a reaction rate. It is argued that smoking causes cardiovascular diseases and some cancers. Many studies have shown associations between smoking level and the incidence of lung cancer, but studies that imply causation in humans have not be carried out.

**Example:** In 1982, SAT scores were first published on a state-by-state basis. Average scores ranged from 790 (out of 1600) (S. Carolina) to 1088 (Iowa). The data were used to support conflicting positions and claims regarding secondary education. In response, Powell and Steelman\textsuperscript{17} attempted to use the data to understand why state-to-state variation was so great and, as they stated, *assess the extent to which the compositional/demographic and school-structural characteristics are implicated in SAT differences.*

These data are sometimes called ecological data. Such data are aggregate statistics—usually means. The data (means) differ from observations on individuals, and so inferences drawn about individuals from these data are potentially misleading or incorrect. In this example, data are measurements on states, not individual high school students, and the reader should not draw inferences about individuals from the data.

Powell and Steelman obtained measurements on a number of variables, but in particular

1. The percentage of total eligible students that took the test

2. The average number of years that the takers took formal courses in social sciences, humanities, and natural sciences

3. The median percentile ranking of the test takers among the students in their classes

The figure below and left is a dotplot showing mean SAT scores by state, and the figure below and right plots rank against percent takers.

Midwestern states tend to have low percentages of test takers, principally because most midwestern state universities required that applicants take the ACT test. Presumably, most of the students that took the SAT in these midwestern states were interested in attending college out-of-state. The percentage of test takers may have a substantial role in explaining variation in state SAT means but little if any insight into the school-structural characteristics responsible for in SAT differences.

It’s expected that median class rank (of the takers) also explains some of the variation in mean state SAT scores. Differences between states with respect to these variables will confound a comparison of states aimed at understanding the success (or lack thereof) of states at educating students. The other potential explanatory variables (income, years, expenditure and public) are far more interesting with respect to the characteristics of states that succeed in educating students.

The plot below and left shows the relationship between SAT means and (mean) years of
formal study. For these data, $r = .330$. If the confounding variables percent takers and median class rank are accounted, or adjusted for, then the apparent relationship is stronger. The figure below and right plots the adjusted mean SAT scores against adjusted years of formal study. For these data, $r = .561$.

Percent takers and median class rank are called confounding variables because the variables affect SAT means and also are associated with years of formal schooling. Differences among states with respect to percent takers and median class rank confounds, or obscures the relationship between mean SAT and mean years of formal schooling. Percent takers and median class rank are confounding variables, and are sometimes called lurking variables particularly if they are not measured, but understood to confound the relationship between two other measured variables.