AP Statistics

Chapter 6

Scatterplots, Association, and Correlation
Objectives:

- Scatterplots
- Association
- Outliers
- Response Variable
- Explanatory Variable
- Correlation
- Correlation Coefficient
- Lurking Variables
Scatterplots

• Shows the relationship between two quantitative variables measured on the same individuals.

• The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.

• Each Individual in the data appears as the point in the plot fixed by the values of both variables for that individual.
Scatterplots usually don’t show the origin (or have breaks), because often neither variable has values near 0. The plot should focus on the part of the coordinate plane that actually contains the data.
Analyzing Scatterplots

Analysis of bivariate data builds on the tools used for examining univariate data.

1. First plot the data, then compute numerical summaries.

2. Look for overall patterns and deviations from those patterns.

3. When the overall pattern is quite regular, use a compact mathematical model to describe it.
Scatterplots

• Scatterplots may be the most common and most effective display for bivariate quantitative data.
  – In a scatterplot, you can see patterns, trends, relationships, and even the occasional extraordinary value sitting apart from the others.
• Scatterplots are the best way to start observing the relationship and the ideal way to picture associations between two quantitative variables.
Scatterplot Variables

• **Response Variable** - measures the outcome of a study (the variable of interest). The variable you hope to predict or explain. (The dependent variable, plotted on the y-axis).

• **Explanatory or Predictor Variable** – helps explain or predict changes in a response variable. Attempts to explain the observed response. (The independent variable, plotted on the x-axis.)
Roles for Variables

• It is important to determine which of the two quantitative variables goes on the $x$-axis and which on the $y$-axis.

• This determination is made based on the roles played by the variables.

• When the roles are clear, the explanatory or predictor variable goes on the $x$-axis, and the response variable (variable of interest) goes on the $y$-axis.
Roles for Variables

• The roles that we choose for variables are more about how we think about them rather than about the variables themselves.

• Just placing a variable on the $x$-axis doesn’t necessarily mean that it explains or predicts anything. And the variable on the $y$-axis may not respond to it in any way.
Example

Here we plotted prediction error on the y-axis against the year on the x-axis because we are interested in how the predictions have changed over time.

Could we have plotted them the other way?

In this case, it is hard to imagine reversing the variables - knowing the prediction error and wanting to guess in what year it happened.
### TABLE 3.1 Average degree-days and natural gas consumption for the Sanchez household

<table>
<thead>
<tr>
<th>Month</th>
<th>Degree-days</th>
<th>Gas (100 cu. ft.)</th>
<th>Month</th>
<th>Degree-days</th>
<th>Gas (100 cu. ft.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov.</td>
<td>24</td>
<td>6.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dec.</td>
<td>51</td>
<td>10.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan.</td>
<td>43</td>
<td>8.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb.</td>
<td>33</td>
<td>7.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mar.</td>
<td>26</td>
<td>5.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apr.</td>
<td>13</td>
<td>4.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>May</td>
<td>4</td>
<td>1.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>June</td>
<td>0</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>July</td>
<td>0</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aug.</td>
<td>1</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sept.</td>
<td>6</td>
<td>2.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct.</td>
<td>12</td>
<td>3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov.</td>
<td>30</td>
<td>6.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dec.</td>
<td>32</td>
<td>7.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan.</td>
<td>52</td>
<td>11.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb.</td>
<td>30</td>
<td>6.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Source: Data provided by Robert Dale, Purdue University.*
Answer:
Association

- The scatterplot displays the association (relationship) between the explanatory variable and the response variable.
- Calling one variable explanatory and the other response doesn’t necessarily mean that changes in one *causes* changes in the other.
- We will discuss causation in more detail later in the chapter.
Looking at Scatterplots

• When looking at scatterplots, we will look for direction, form, strength, and unusual features.

• Direction:
  – A pattern that runs from the upper left to the lower right is said to have a **negative** direction.
  – A trend running the other way has a **positive** direction.
Direction

Positive
• Same as a positive slope, a direct association.

Negative
• Same as a negative slope, an inverse association.
Looking at Scatterplots

- This example shows a negative association between central pressure and maximum wind speed.
- As the central pressure increases, the maximum wind speed decreases.
Can the NOAA predict where a hurricane will go?

- The figure shows a negative direction between the years since 1970 and the prediction errors made by NOAA.
- As the years have passed, the predictions have improved (errors have decreased).
Form

- Type of pattern, such as linear or nonlinear.
Looking at Scatterplots

- **Form:**
  - If there is a straight line (linear) relationship, it will appear as a cloud or swarm of points stretched out in a generally consistent, straight form.
• Form:
  – If the relationship isn’t straight, but curves gently, while still increasing or decreasing steadily, we can often find ways to make it more nearly straight.
Looking at Scatterplots

• Form:
  – If the relationship curves sharply, the methods of this book cannot really help us.
Strength

- How closely the points follow a clear form. (none, weak, moderate, or strong)
Looking at Scatterplots

• Strength:
  – At one extreme, the points appear to follow a single stream

(whether straight, curved, or bending all over the place).
Looking at Scatterplots

• Strength:
  – At the other extreme, the points appear as a vague cloud with no discernable trend or pattern:
  – Note: we will quantify the amount of scatter soon.
Examples: Direction, Form, and Strength

Positive, linear, moderate association

Positive, linear, strong association

Positive, linear, perfect association

Negative, linear, weak association

Negative, linear, strong association

Negative, linear, perfect association

No association

Non-linear association
Looking at Scatterplots

- Unusual features:
  - Look for the unexpected.
  - Often the most interesting thing to see in a scatterplot is the thing you never thought to look for.
  - One example of such a surprise is an outlier standing away from the overall pattern of the scatterplot.
  - Clusters or subgroups should also raise questions.
Outliers

• An important unusual feature is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

• As before, outliers are almost always interesting and always deserve special attention.

• You should also look for clusters or subgroups that deviate from the pattern. Deviating groups raise questions about why they are different. They may indicate that you should split the data into different groups, instead of looking at it all together.
Categorical Variables in Scatterplots

• Adding a categorical variable to a scatterplot introduces a third variable to the scatterplot.
• This categorical variable has only two values.
• The two values are displayed by two different plotting symbols or colors.
Purpose of Adding a Categorical Variable

- Can be used to look for deviations from the pattern, such as outliers, and other internal patterns.
Example – Scatterplot with an added categorical variable
Another Example
# Problem

## Table 3.2: Lean body mass and metabolic rate

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sex</th>
<th>Mass (kg)</th>
<th>Rate (cal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>62.0</td>
<td>1792</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>62.9</td>
<td>1666</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>36.1</td>
<td>995</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>54.6</td>
<td>1425</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>48.5</td>
<td>1396</td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>42.0</td>
<td>1418</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
<td>47.4</td>
<td>1362</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>50.6</td>
<td>1502</td>
</tr>
<tr>
<td>9</td>
<td>F</td>
<td>42.0</td>
<td>1256</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>48.7</td>
<td>1614</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sex</th>
<th>Mass (kg)</th>
<th>Rate (cal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>F</td>
<td>40.3</td>
<td>1189</td>
</tr>
<tr>
<td>12</td>
<td>F</td>
<td>33.1</td>
<td>913</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>51.9</td>
<td>1460</td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>42.4</td>
<td>1124</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>34.5</td>
<td>1052</td>
</tr>
<tr>
<td>16</td>
<td>F</td>
<td>51.1</td>
<td>1347</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>41.2</td>
<td>1204</td>
</tr>
<tr>
<td>18</td>
<td>M</td>
<td>51.9</td>
<td>1867</td>
</tr>
<tr>
<td>19</td>
<td>M</td>
<td>46.9</td>
<td>1439</td>
</tr>
</tbody>
</table>
Correlation

• Exists between two variables when one of them is related to the other in some way.
• Our eyes are not good judges of how strong a linear relationship is.
• Compare the strength of the following two scatterplots, which has the stronger association?
B.
They are the **SAME**, with different scales
Therefore, we have a numerical measure to analyze the strength (and direction) of a scatterplot.
Correlation

• Data collected from students in Statistics classes included their heights (in inches) and weights (in pounds):

• Here we see a positive association and a fairly straight form, although there seems to be a high outlier.
Correlation

- How strong is the association between weight and height of Statistics students?
- If we had to put a number on the strength, we would not want it to depend on the units we used.
- Changing the units, a scatterplot of heights (in centimeters) and weights (in kilograms) doesn’t change the shape of the pattern:
• Since the units don’t matter, why not remove them altogether?
• We could standardize both variables, 
\[
(z_x = \frac{x - \bar{x}}{s_x} \quad \text{and} \quad z_y = \frac{y - \bar{y}}{s_y})
\]
and write the coordinates of a point as \((z_x, z_y)\).
• Here is a scatterplot of the standardized weights and heights:
Correlation

- Note that the underlying linear pattern seems steeper in the standardized plot than in the original scatterplot.
- That’s because we made the scales of the axes the same.
- Equal scaling gives a neutral way of drawing the scatterplot and a fairer impression of the strength of the association.
Correlation

• Which points in the scatterplot of z-scores give the impression of positive association?
• In a positive association, $y$ increases as $x$ increases.
• Some points (those in green) strengthen the impression of a positive association between height and weight.
• For these points, $z_x$ and $z_y$, have the same sign, so the product $z_xz_y$ is positive.
Correlation

- The points in red tend to weaken the positive association (or support a negative association).
- For these points, $z_x$ and $z_y$ have opposite signs. So the product $z_x z_y$ is negative.
Correlation

• To turn these products into a measure of strength of association, just add up the $z_x z_y$ products for every point in the scatterplot.

$$\sum z_x z_y$$

• This summarizes the direction and strength of association for all the points.

• If most of the points are green, the sum will tend to be positive.

• If most of the points are red, the sum will tend to be negative.
Correlation

- But the size of this sum gets bigger the more data we have.
- So, to adjust for this, we divide the sum by \( n-1 \).
- This ratio; \[ \frac{\sum z_x z_y}{n - 1} \]
  is the correlation coefficient.
Correlation

• For the students’ heights and weights, the correlation is 0.644.
• What does this mean in terms of strength? We’ll address this shortly.
Correlation

• The correlation coefficient \((r)\) gives us a numerical measurement of the strength of the linear relationship between the explanatory and response variables.

\[
r = \frac{\sum z_x z_y}{n - 1}
\]
Correlation

• Therefore, the correlation ($r$) is an average of the products of the standardized $x$ variable and the standardized $y$ variable for $n$ individuals.

• Calculate correlation ($r$) on the TI-83/84, automatic when calculating LSRL, must have DiagnosticOn under CATALOG function. Will discuss further in Chapter 8, Linear Regression.
Correlation

- Measures the strength and direction of the **LINEAR** relationship between two quantitative variables.
- Data required on variables $x$ & $y$ for $n$ individuals → means $(\bar{x}, \bar{y})$ and standard deviations $(S_x, S_y)$.

Correlation

$$ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{S_x} \right) \left( \frac{y_i - \bar{y}}{S_y} \right) $$
Finding the correlation coefficient by hand

To find the correlation coefficient by hand, start with the summary statistics for both variables: $\bar{x}$, $\bar{y}$, $s_x$, and $s_y$. Then find the deviations as we did for the standard deviation, but now in both $x$ and $y$: $(x - \bar{x})$ and $(y - \bar{y})$. For each data pair, multiply these deviations together: $(x - \bar{x})(y - \bar{y})$. Add the products up for all data pairs. Finally, divide the sum by the product of $(n - 1) \times s_x \times s_y$ to get the correlation coefficient.

Here we go:

Suppose the data pairs are:

<table>
<thead>
<tr>
<th>$x$</th>
<th>6</th>
<th>10</th>
<th>14</th>
<th>19</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

Then $\bar{x} = 14$, $\bar{y} = 7$, $s_x = 6.20$, and $s_y = 3.39$

<table>
<thead>
<tr>
<th>Deviations in $x$</th>
<th>Deviations in $y$</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 - 14 = -8</td>
<td>5 - 7 = -2</td>
<td>-8 \times -2 = 16</td>
</tr>
<tr>
<td>10 - 14 = -4</td>
<td>3 - 7 = -4</td>
<td>16</td>
</tr>
<tr>
<td>14 - 14 = 0</td>
<td>7 - 7 = 0</td>
<td>0</td>
</tr>
<tr>
<td>19 - 14 = 5</td>
<td>8 - 7 = 1</td>
<td>5</td>
</tr>
<tr>
<td>21 - 14 = 7</td>
<td>12 - 7 = 5</td>
<td>35</td>
</tr>
</tbody>
</table>

Add the products up: $16 + 16 + 0 + 5 + 35 = 72$

Finally, we divide by $(n - 1) \times s_x \times s_y = (5 - 1) \times 6.20 \times 3.39 = 84.07$

The ratio is the correlation coefficient:

$r = 72/84.07 = 0.856$. 
Correlation Conditions

• **Correlation** measures the strength of the *linear* association between two *quantitative* variables.

• Before you use correlation, you must check several conditions:
  – Quantitative Variables Condition
  – Straight Enough Condition
  – Outlier Condition
CorrelationConditions

- Quantitative Variables Condition:
  - Correlation applies only to quantitative variables.
  - Don’t apply correlation to categorical data masquerading as quantitative.
  - Check that you know the variables’ units and what they measure.
Correlation Conditions

• Straight Enough Condition:
  – You can *calculate* a correlation coefficient for any pair of variables.
  – But correlation measures the strength only of the *linear* association, and will be misleading if the relationship is not linear.
Correlation Conditions

• **Outlier Condition:**
  – Outliers can distort the correlation dramatically.
  – An outlier can make an otherwise small correlation look big or hide a large correlation.
  – It can even give an otherwise positive association a negative correlation coefficient (and vice versa).
  – When you see an outlier, it’s often a good idea to report the correlations with and without the point.
Correlation Properties

• The sign of a correlation coefficient gives the direction of the association.
• Correlation is always between $-1$ and $+1$.
  – Correlation *can* be exactly equal to $-1$ or $+1$, but these values are unusual in real data because they mean that all the data points fall *exactly* on a single straight line.
  – A correlation near zero corresponds to no linear association.
Correlation Properties

- $-1 \leq r \leq 1$: The magnitude of $r$ reflects the strength of the linear association as viewed in a scatterplot. An $r$-value of $-1$ represents perfect negative correlation; $r = 1$ represents perfect positive correlation; and $r = 0$ represents no correlation. ($0 \leq |r| < .25$ none, $.25 \leq |r| < .5$ weak, $.5 \leq |r| < .75$ moderate, $.75 \leq |r| < 1$ strong)
Examples Correlation:

(a) $r = 1$
perfect positive correlation

(b) $r = 0.8$
strong positive correlation

(c) $r = 0.3$
moderate to weak positive correlation

(d) $r = 0$
no correlation

(e) $r = -0.3$
moderate to weak negative correlation

(f) $r = -0.8$
strong negative correlation

(g) $r = -1$
perfect negative correlation
More Examples
Correlation:

- Correlation $r = 0$
- Correlation $r = -0.3$
- Correlation $r = 0.5$
- Correlation $r = -0.7$
- Correlation $r = 0.9$
- Correlation $r = -0.99$
Correlation Properties

• Correlation treats $x$ and $y$ symmetrically:
  – The correlation of $x$ with $y$ is the same as the correlation of $y$ with $x$ (no distinction between explanatory and response variables).

• Correlation has no units (uses standardized values).

• Correlation is not affected by changes in the center or scale of either variable.
  – Correlation depends only on the $z$-scores, and they are unaffected by changes in center or scale.
Correlation Properties

- Correlation measures the strength of the *linear* association between the two variables.
  - Variables can have a strong association but still have a small correlation if the association isn’t linear.
- Correlation is sensitive to outliers. A single outlying value can make a small correlation large or make a large one small.
Correlation ≠ Causation

- Whenever we have a strong correlation, it is tempting to explain it by imagining that the predictor variable has caused the response to help.
- Scatterplots and correlation coefficients never prove causation.
- A hidden variable that stands behind a relationship and determines it by simultaneously affecting the other two variables is called a lurking variable.
Correlation ≠ Causation

- **Lurking variable**: A variable that is not explicitly part of a study but affects the way the variables in the study appear to be related.
- A variable, usually unobserved, that influences the association between the variables of primary interest.
- Because we can never be certain that observational data are not hiding a lurking variable, it is never safe to conclude that a scatterplot demonstrates a cause-and-effect relationship, no matter how strong the correlation.
Correlation ≠ Causation

• Scatterplots and correlation coefficients *never* prove causation.

• That’s the reason it took so long for the US Surgeon General to get warning labels on cigarettes. Although there was plenty of evidence that increased smoking was *associated* with lung cancer, it took years to provide evidence that smoking actually *causes* lung cancer.

• Because it is so important, we will repeat it: A scatterplot or correlation alone *cannot* prove causation.
Correlation Tables

It is common in some fields to compute the correlations between each pair of variables in a collection of variables and arrange these correlations in a table.

<table>
<thead>
<tr>
<th></th>
<th>Assets</th>
<th>Sales</th>
<th>Market Value</th>
<th>Profits</th>
<th>Cash Flow</th>
<th>Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assets</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sales</td>
<td>0.746</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market Value</td>
<td>0.682</td>
<td>0.879</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profits</td>
<td>0.602</td>
<td>0.814</td>
<td>0.968</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cash Flow</td>
<td>0.641</td>
<td>0.855</td>
<td>0.970</td>
<td>0.989</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Employees</td>
<td>0.594</td>
<td>0.924</td>
<td>0.818</td>
<td>0.762</td>
<td>0.787</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Correlation Tables

- Correlation tables are compact and give a lot of summary information at a glance. They can be an efficient way to look at large data sets, but also a dangerous one.

- By presenting all these correlations without any checks for linearity and outliers, the correlation table risks showing misleading or meaningless information due to hidden outliers or a nonlinear form.
Straightening Scatterplots

- Straight line relationships are the ones that we can measure with correlation.
- When a scatterplot shows a bent form that consistently increases or decreases, we can often straighten the form of the plot by re-expressing one or both variables.
Straightening Scatterplots

• A scatterplot of f/stop vs. shutter speed shows a bent relationship:

![Scatterplot diagram](image)
Re-expressing f/stop vs. shutter speed by squaring the f/stop values straightens the relationship.
What Can Go Wrong?

• Don’t say “correlation” when you mean “association.”
  – More often than not, people say correlation when they mean association.
  – The word “correlation” should be reserved for measuring the strength and direction of the linear relationship between two quantitative variables.
What Can Go Wrong?

• Don’t correlate categorical variables.
  – Be sure to check the Quantitative Variables Condition.

• Don’t confuse “correlation” with “causation.”
  – Scatterplots and correlations never demonstrate causation.
  – These statistical tools can only demonstrate an association between variables.
What Can Go Wrong?

- Be sure the association is linear.
  - There may be a strong association between two variables that have a nonlinear association.
What Can Go Wrong?

- Don’t assume the relationship is linear just because the correlation coefficient is high.

- Here the correlation is 0.979, but the relationship is actually bent.
What Can Go Wrong?

- Beware of outliers.
  - Even a single outlier can dominate the correlation value.
  - Make sure to check the Outlier Condition.
What have we learned?

- We examine scatterplots for *direction, form, strength,* and *unusual features.*
- Although not every relationship is linear, when the scatterplot is straight enough, the *correlation coefficient* is a useful numerical summary.
  - The sign of the correlation tells us the direction of the association.
  - The magnitude of the correlation tells us the *strength* of a linear association.
  - Correlation has no units, so shifting or scaling the data, standardizing, or swapping the variables has no effect on the numerical value.
What have we learned?

• Doing Statistics right means that we have to *Think* about whether our choice of methods is appropriate.
  – Before finding or talking about a correlation, check the Straight Enough Condition.
  – Watch out for outliers!

• Don’t assume that a high correlation or strong association is evidence of a cause-and-effect relationship—beware of lurking variables!
Assignment

- Exercises pg. 164 – 170: #3, 5, 6, 11, 12, 14, 15, 16, 20, 23, 27, 28
- Read Ch-8, pg. 171 - 192