How can we understand the relationship between two variables?

Thought Question

http://www.tylervigen.com/

How strong is the linear relationship between per capita consumption of mozzarella cheese and the number of civil engineering doctorates awarded?

Do you think an increase in the per capita consumption of mozzarella cheese causes an increase in the number of civil engineering doctorates awarded, or vice versa?
How strong is the linear relationship between per capita consumption of mozzarella cheese and the per capita consumption of cheese?

Do you think an increase in the per capita consumption of mozzarella cheese causes an increase in the per capita consumption of cheese?
Regression Lines

How can we summarize the relationship between an explanatory variable and a response variable?

Terminology

- **Regression line**: straight line that describes how a response variable $y$ changes as an explanatory variable $x$ changes.

Example: Botulism

The following data refer to an outbreak of botulism. Each case is a person who died from botulism in the outbreak. The variables recorded are the subject’s age (in years) and the incubation period (in hours).

<table>
<thead>
<tr>
<th>Case</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>29</td>
<td>37</td>
<td>42</td>
<td>38</td>
<td>51</td>
<td>30</td>
<td>32</td>
<td>33</td>
<td>31</td>
<td>32</td>
<td>50</td>
</tr>
<tr>
<td><strong>Incubation</strong></td>
<td>13</td>
<td>34</td>
<td>20</td>
<td>18</td>
<td>19</td>
<td>36</td>
<td>48</td>
<td>21</td>
<td>32</td>
<td>48</td>
<td>16</td>
</tr>
</tbody>
</table>

The following plot is a scatterplot of this data along with a regression line.
What is the explanatory variable?

age

What is the response variable?

incubation period

If we use the regression line to predict the incubation period for other ages, how accurate do you think our predictions will be?

Not very accurate, but not completely off base.
Regression Equations

How do we determine the “best” regression line?

Terminology

- **Least-squares regression line**: the line that makes the sum of squares of the vertical distances of the data points from the line as small as possible.

Example: Archaeopteryx Bones

The least-squares line is

$$\text{humerus length} = -3.66 + 1.197 \times \text{femur length}.$$
What is the slope of the line? What is its interpretation?

The slope is 1.197. When femur length increases by 1 cm, humerus length increases by 1.197 cm.

What is the intercept of the line? Can we interpret it meaningfully?

The intercept is -3.66. This is the value of the predicted humerus length when the femur length is 0 cm. However, the femur cannot have length 0 cm.

Use the least-squares line to predict the humerus length for a fossil with a femur 70 cm long.

humerus length = -3.66 + 1.197 \times 70 = 80.13 \text{ cm}.
Properties of Prediction

- Prediction is based on fitting some “model” to a set of data.

- Prediction works best when the model fits the data closely.

- Prediction outside the range of the available data is risky.

Example: Height and Age

The following regression equation relates Sarah’s age (in months) to her height (in cm), using data collected until Sarah was 5 years (60 months) old:

\[ \text{height} = 71.95 + 0.383 \times \text{age} \]

Approximately how tall was Sarah when she was 42 months old?

height = 71.95 + 0.383 \times 42 \approx 88 \text{ cm.}

So Sarah was about 3 feet tall when she was 42 months old.

Use the regression equation to predict Sarah’s height at 30 years old.

height = 71.95 + 0.383 \times 360 \approx 210 \text{ cm.}

So we predict Sarah will be about 6 feet 10 inches when she is 30 years old.

Is this prediction reasonable?

No!
Correlation and Regression

What is the relationship between the correlation $r$ and a regression line?

**Similarities and Differences**

- Both are statistical techniques that use data for two quantitative variables.

- The correlation $r$ measures the strength and direction of a straight-line relationship with a single number; e.g., $r = 0.84$.

- A regression line is a mathematical equation that describes the relationship.

- Both are strongly affected by outliers.

**Connection**

The usefulness of the regression line for prediction depends on the strength of the straight-line association (which is measured by the correlation).

In fact, the square of the correlation is the right measure of the strength of the association.

**Terminology**

- **Square of the correlation** $r^2$: the proportion of the variation in the values of $y$ that is explained by the least-squares regression of $y$ on $x$. 
Example: Archaeopteryx Bones

The correlation between the femur length and the humerus length is $r = 0.994$.

*Calculate $r^2$.*

$$r^2 = (0.994)^2 = 0.988.$$ 

*Interpret this result in terms of proportions of variation.*

Variation in femur length accounts for 98.8% of all the variation in humerus length. Other factors (resulting in the scatter of the points about the line) accounts for the remaining 1.2%.

*How accurate do you think predicted values of the humerus length will be?*

Very accurate.
Causation

If two variables are correlated, does one cause the other?

**Observations**

- Even very strong correlations may not correspond to a real causal relationship between two variables.

- The relationship between two variables is often influenced by other variables lurking in the background.

- Even if the explanatory variable causes some change in the response variable, is rarely the only influence on the response variable.

- The best evidence for causation comes from randomized comparative experiments.

Correlation does not always equal causation!

http://www.xkcd.com/
Example: Fires

A Chicago newspaper reported that “there is a strong correlation between the number of fire trucks at a fire and the amount of damage that the fire does.” There is a strong linear relationship between the number of fire trucks and the amount of damage.

Do you think that the increase in the number of fire trucks actually causes the increase in damage?

No

What lurking variable influences both the number of fire trucks and the increase in damage?

The size and strength of the fire.

Reasons Why Two Variables May Be Related

- The explanatory variable causes a change in the response variable (direct causation).

- The observed association is due to a lurking variable that affects both the explanatory variable and the response variable (common response).

- Both the explanatory variable and a lurking variable influence the response variable. We can’t distinguish their effects on the response variable (confounding).

- The observed relationship could be due to two or more of direct causation, common response, or confounding.

- The correlation may be merely a coincidence!
Examples

What explains the relationship between per capita consumption of mozzarella cheese and the number of civil engineering doctorates awarded?

Coincidence.

What explains the relationship between per capita consumption of mozzarella cheese and per capita consumption of cheese?

There is a direct causation, because consuming more mozzarella cheese means more cheese has been consumed overall, but there is also confounding because other factors can affect how much cheese is consumed overall.

What explains the relationship between the number of fire trucks and the increase in damage?

Common response - the size of the fire affects both variables.
Evidence for Causation

How do we establish causation when we can’t do an experiment?

Criteria

- The association is strong.
- The association is consistent.
- Higher doses are associated with stronger results.
- The alleged cause precedes the effect in time.
- The alleged cause is plausible.
Chapter 15 Exercises

1. A group of researchers believe that drinking wine regularly can help prevent heart disease. They measure yearly wine consumption (liters of alcohol from drinking wine, per person) and yearly deaths from heart disease (deaths per 100,000 people) for 6 countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Liters of Alcohol</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>1.5</td>
<td>80</td>
</tr>
<tr>
<td>France</td>
<td>8.5</td>
<td>40</td>
</tr>
<tr>
<td>Italy</td>
<td>7.5</td>
<td>60</td>
</tr>
<tr>
<td>Spain</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>2.75</td>
<td>120</td>
</tr>
<tr>
<td>United States</td>
<td>1.25</td>
<td>120</td>
</tr>
</tbody>
</table>

The correlation is \( r = -0.8079 \), and the equation of the least-squares regression line is

\[
\text{death rate} = 118.576 - 9.111 \times \text{liters of alcohol}.
\]

(a) Make a scatterplot that shows the relationship between wine consumption and deaths from heart disease. Sketch the regression line on your scatterplot.

(b) What do the scatterplot and the correlation tell you about the direction and strength of the association between these two variables?

The association is negative and quite strong.
(c) Use the least-squares regression line to predict the heart disease death rate for a country in which the yearly wine consumption is 2 liters per person.

\[
\text{death rate} = 118.576 - 9.111 \times 2 \approx 100 \text{ deaths per 100,00 people.}
\]

(c) Can we conclude that drinking wine regularly can help prevent heart disease? What are some differences in the countries that may be confounded with wine-drinking habits?

Not necessarily - there may be some lurking variables confounding the results, such as diet and genetic background.