Chapter 2

Looking at Data—Relationships

Introduction to the Practice of

STATISTICS EIGHTH EDITION

Moore / McCabe / Craig

Lecture Presentation Slides

Chapter 2 Looking at Data—Relationships

2.1 Relationships

2.2 Scatterplots

2.3 Correlation

2.4 Least-Squares Regression

2.5 Cautions about Correlation and Regression

2.6 Data Analysis for Two-Way Tables

2.7 The Question of Causation

2.1 Relationships

- What is an association between variables?
- Explanatory and response variables
- Key characteristics of a data set

Associations Between Variables

Many interesting examples of the use of statistics involve relationships between pairs of variables.

> Two variables measured on the same cases are **associated** if knowing the value of one of the variables tells you something that you would not otherwise know about the value of the other variable.

When you examine the relationship between two variables, a new question becomes important: *Is your purpose simply to explore the nature of the relationship, or do you wish to show that one of the variables can explain variation in the other?*

> A **response variable** measures an outcome of a study. An **explanatory variable** explains or causes changes in the response variable.

Key Characteristics of a Data Set

Certain characteristics of a data set are key to exploring the relationship between two variables. These should include the following:

- ✓ Cases: Identify the cases and how many there are in the data set.
- ✓ Label: Identify what is used as a label variable if one is present.
- ✓ Categorical or quantitative: Classify each variable as categorical or quantitative.
- ✓ Values: Identify the possible values for each variable.
- ✓ Explanatory or response: If appropriate, classify each variable as explanatory or response.

2.2 Scatterplots

- Scatterplots
- Interpreting scatterplots
- Categorical variables in scatterplots

Scatterplot

The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot.**

A **scatterplot** 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 corresponds to one point on the graph.

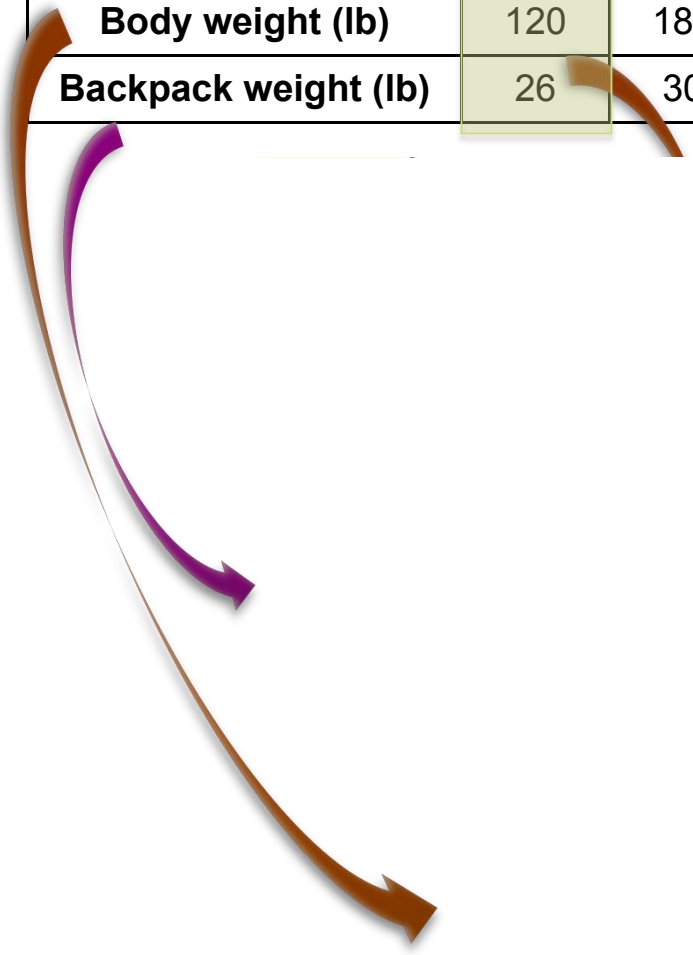
How to Make a Scatterplot

1. Decide which variable should go on each axis. If a distinction exists, plot the explanatory variable on the *x* axis and the response variable on the *y* axis.
2. Label and scale your axes.
3. Plot individual data values.

Scatterplot

Example: Make a scatterplot of the relationship between body weight and backpack weight for a group of hikers.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28



Interpreting Scatterplots

To interpret a scatterplot, follow the basic strategy of data analysis from Chapter 1. Look for patterns and important departures from those patterns.

How to Examine a Scatterplot

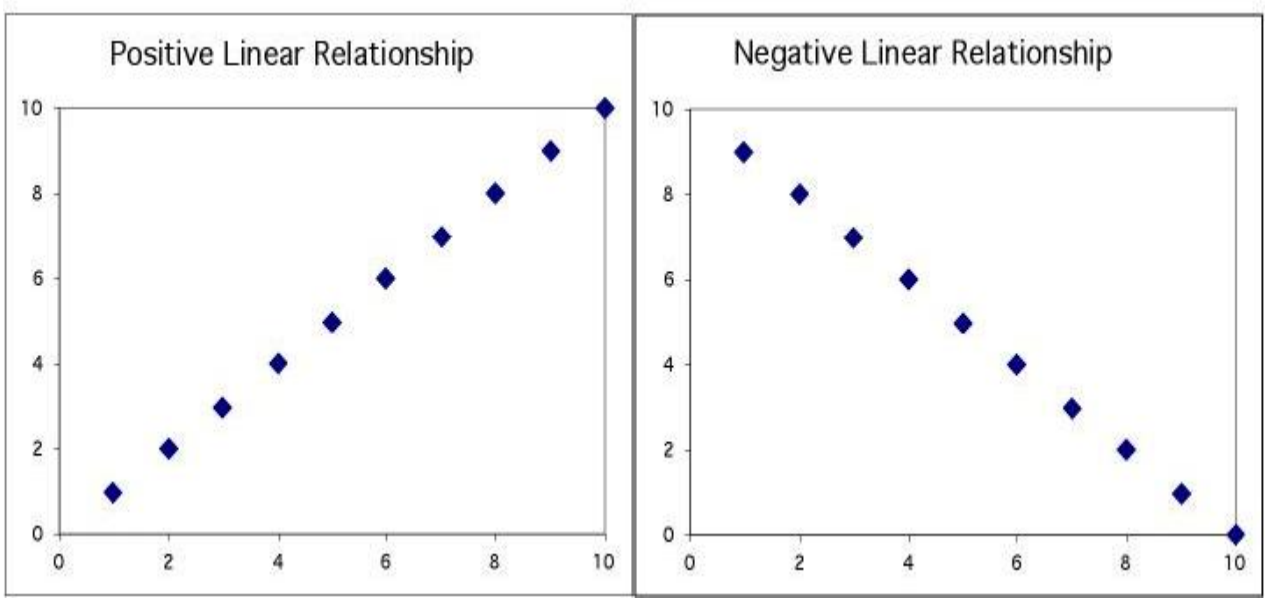
As in any graph of data, look for the *overall pattern* and for striking *deviations* from that pattern.

- You can describe the overall pattern of a scatterplot by the **direction, form,** and **strength** of the relationship.
- An important kind of departure is an **outlier,** an individual value that falls outside the overall pattern of the relationship.

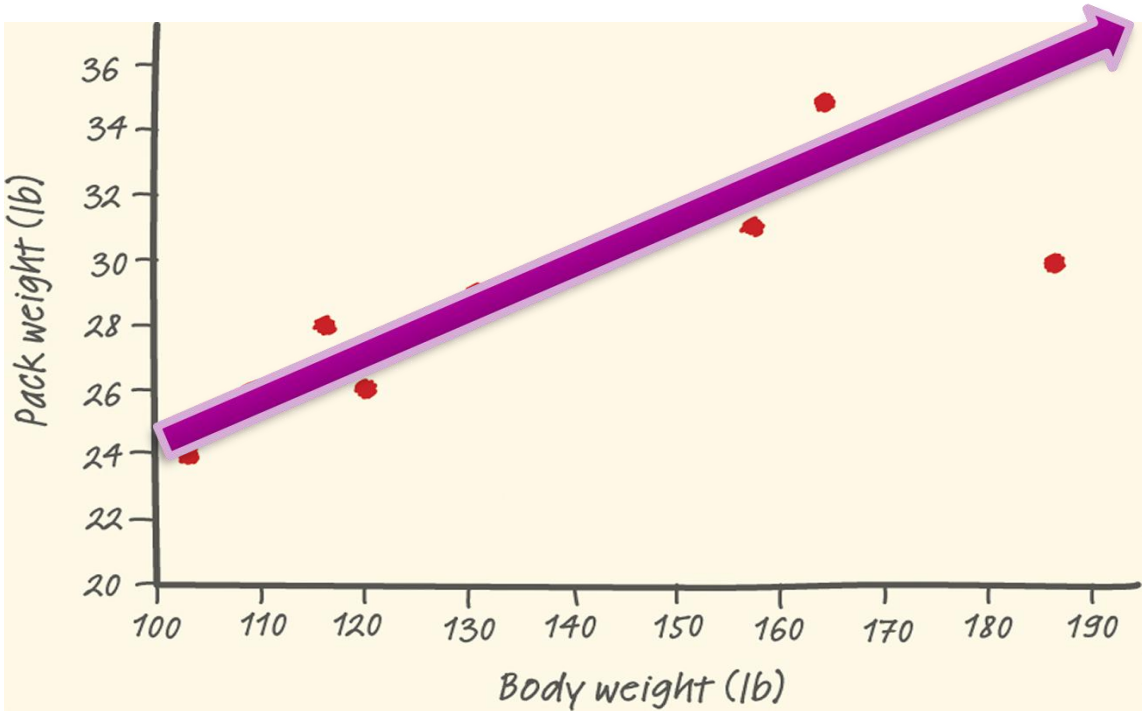
Interpreting Scatterplots

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice-versa.



Interpreting Scatterplots



Outlier

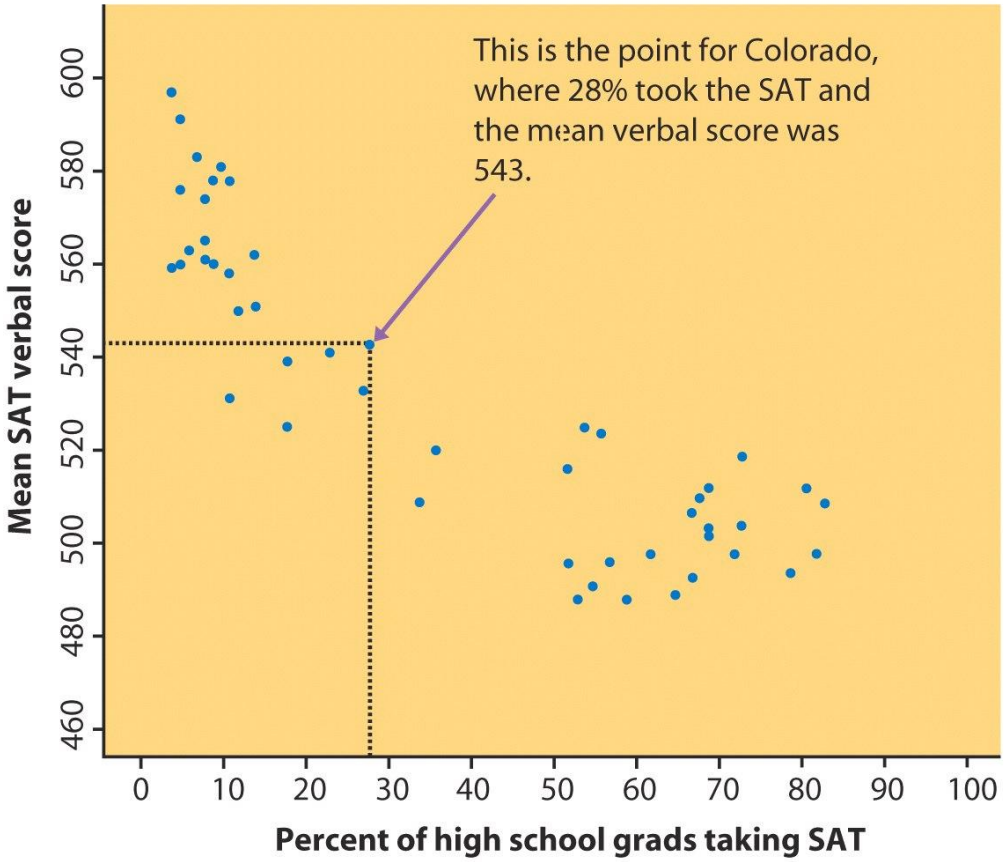
- ✓ There is one possible outlier—the hiker with the body weight of 187 pounds seems to be carrying relatively less weight than are the other group members.

Strength **Direction** **Form**

- ✓ There is a moderately strong, positive, linear relationship between body weight and backpack weight.
- ✓ It appears that lighter hikers are carrying lighter backpacks.

Adding Categorical Variables

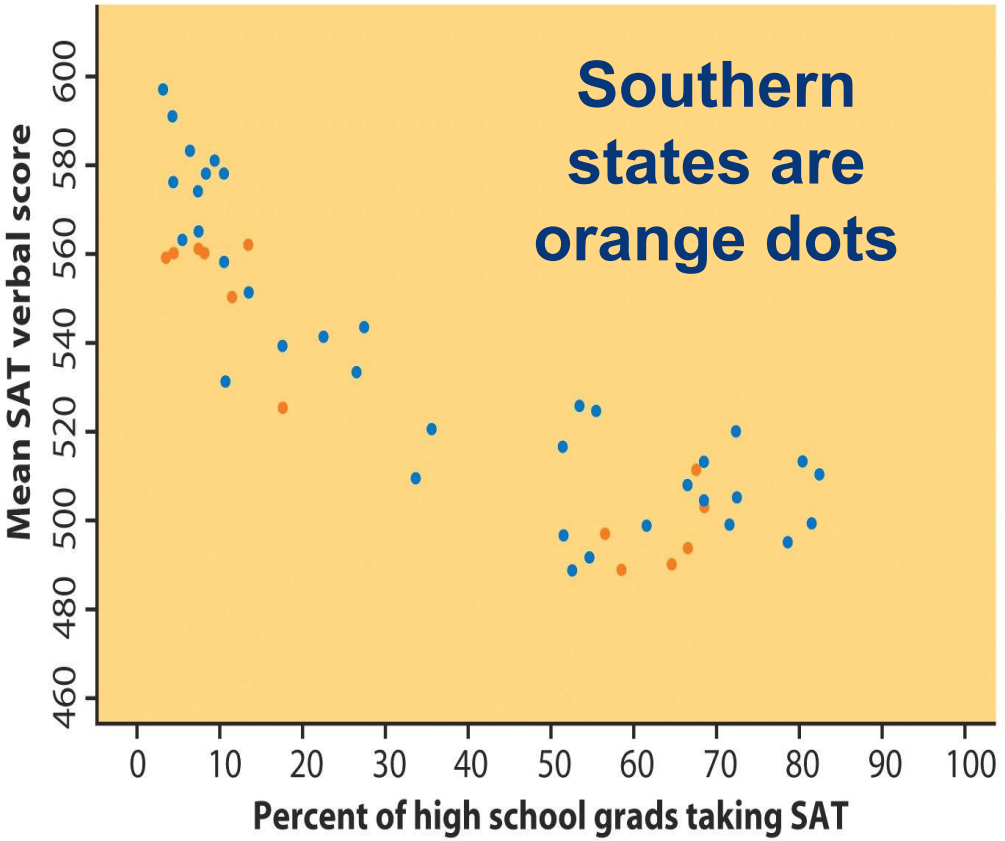
- Consider the relationship between mean SAT verbal score and percent of high school grads taking the SAT for each state.



Adding Categorical Variables

- Consider the relationship between mean SAT verbal score and percent of high school grads taking the SAT for each state.

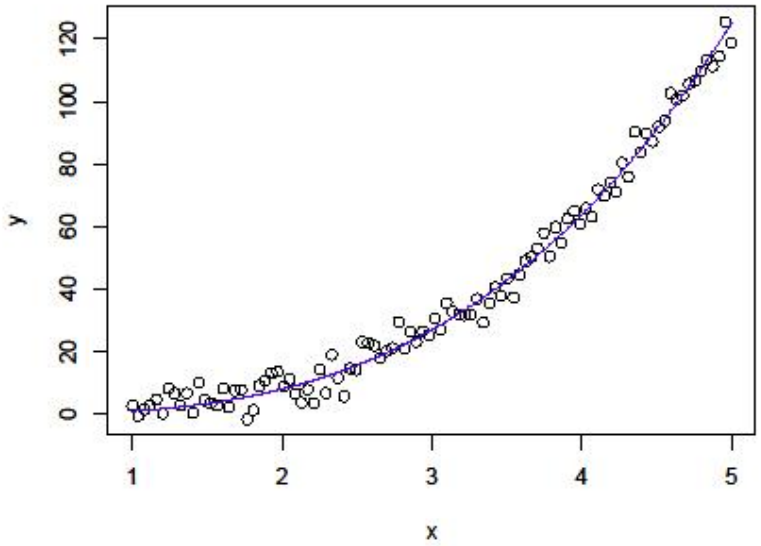
To add a *categorical variable*, use a different plot color or symbol for each category.



Nonlinear Relationships

- There are other forms of relationships besides linear. The scatterplot below is an example of a nonlinear form.

- Note that there is curvature in the relationship between *x* and *y*.



2.3 Correlation

- The correlation coefficient *r*
- Properties of *r*
- Influential points

Measuring Linear Association

A scatterplot displays the strength, direction, and form of the relationship between two quantitative variables. Linear relationships are important because a straight line is a simple pattern that is quite common.

Our eyes are not always good judges of how strong a relationship is. Therefore, we use a numerical measure to supplement our scatterplot and help us interpret the strength of the linear relationship.

The **correlation *r*** measures the strength of the linear relationship between two quantitative variables. Using the notation explained on pp. 103–104 in the text:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

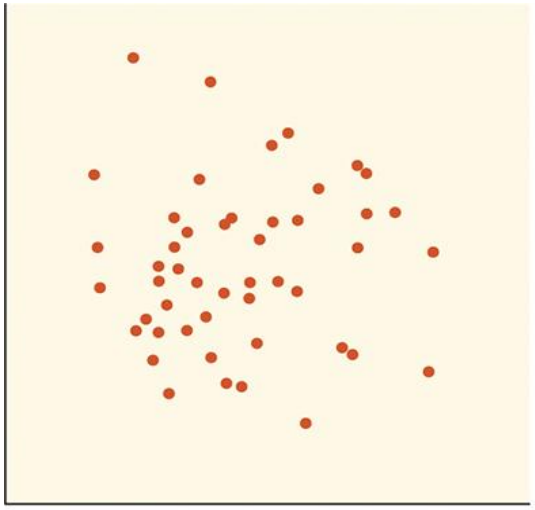
Measuring Linear Association

We say a linear relationship is strong if the points lie close to a straight line and weak if they are widely scattered about a line. The following facts about *r* help us further interpret the strength of the linear relationship.

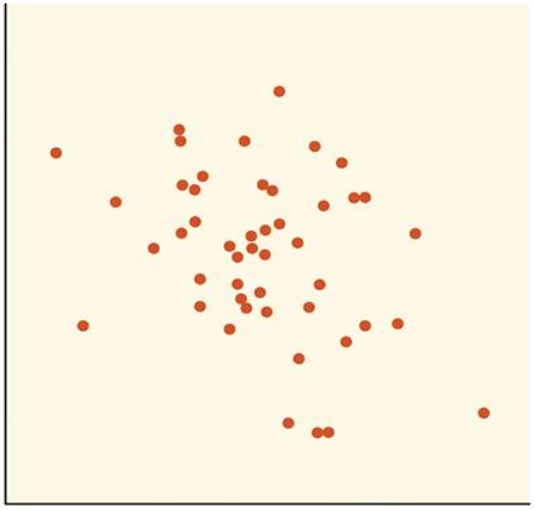
Properties of Correlation

- *r* is always a number between –1 and 1.
- $r > 0$ indicates a positive association.
- $r < 0$ indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 toward –1 or 1.
- The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship.

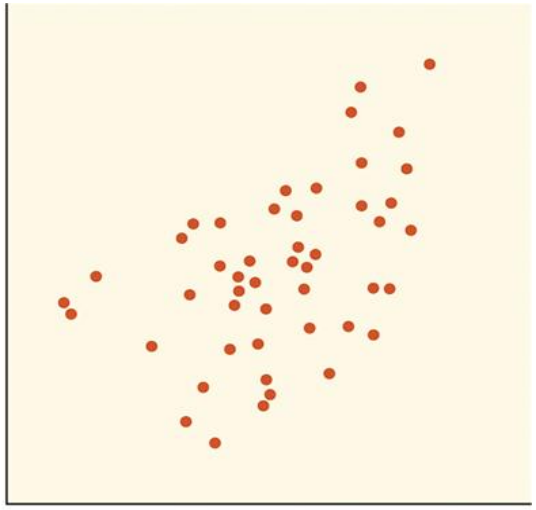
Correlation



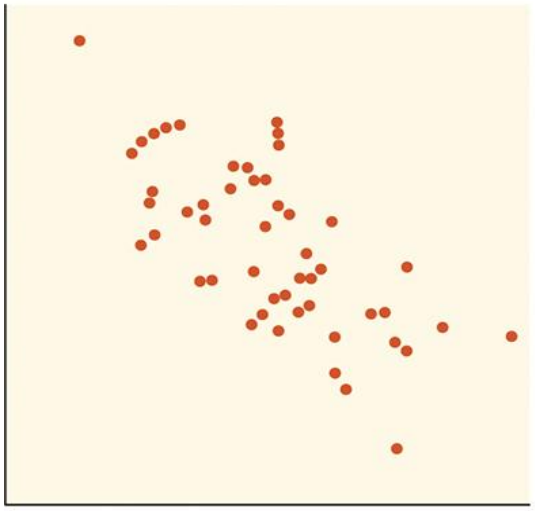
Correlation $r = 0$



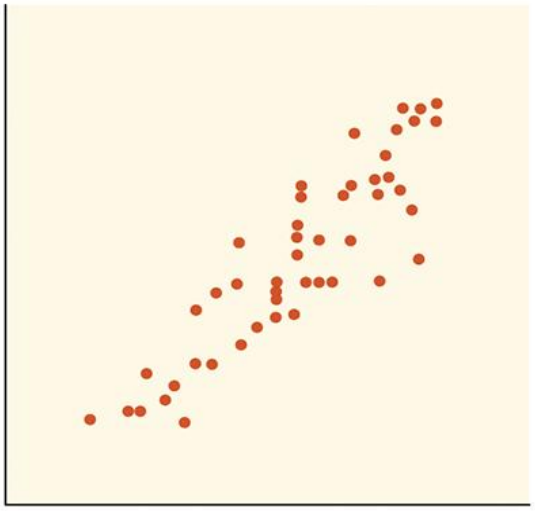
Correlation $r = -0.3$



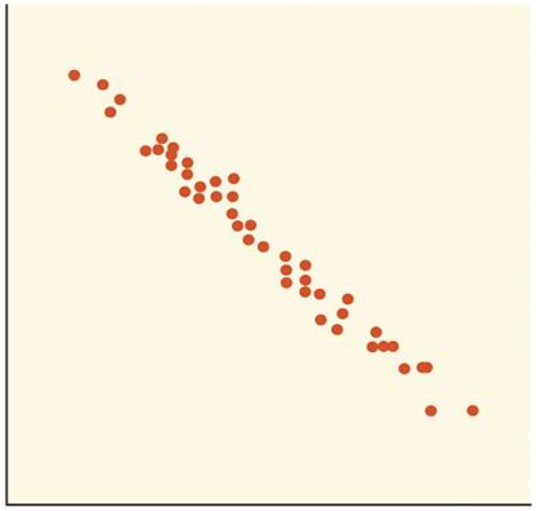
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

Properties of Correlation

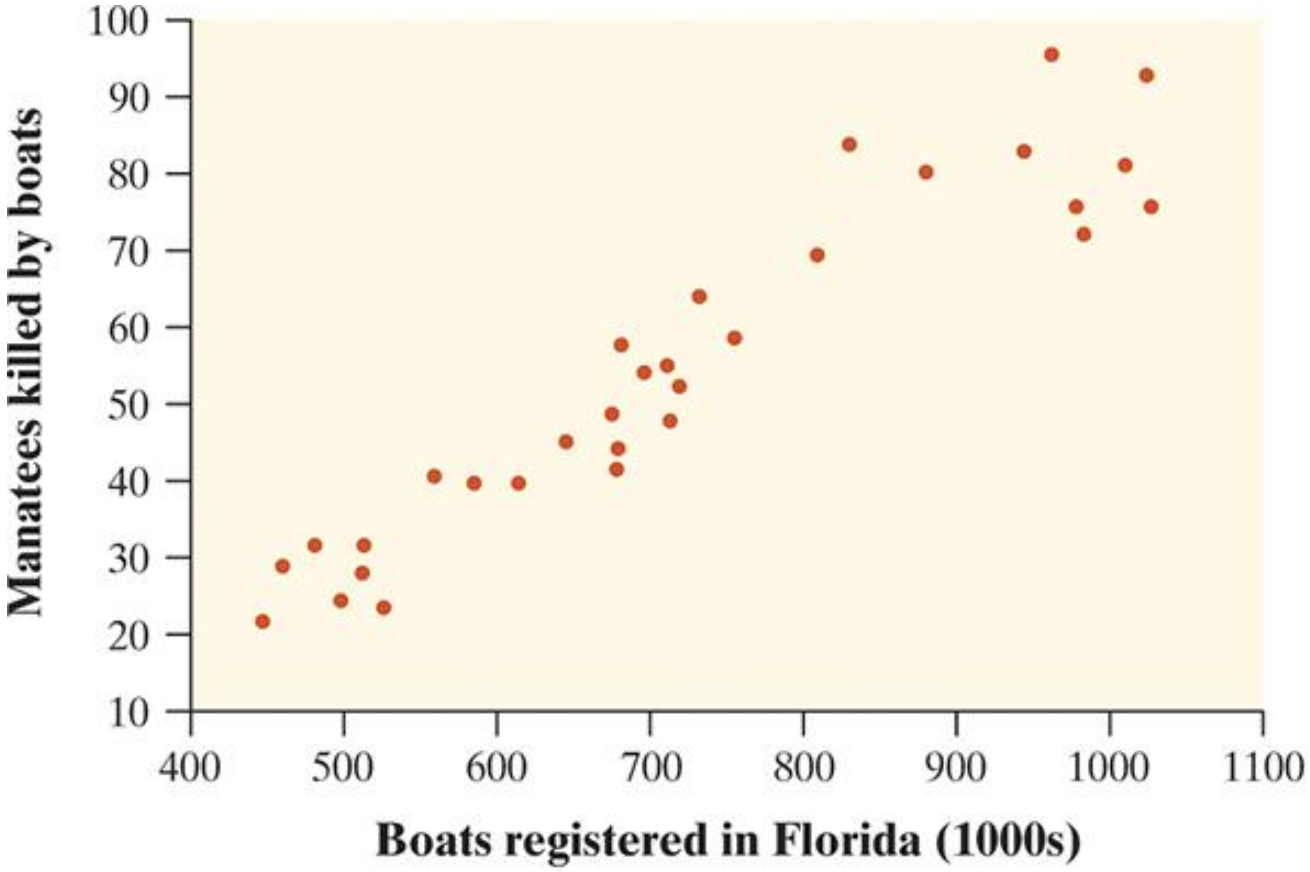
1. Correlation makes no distinction between explanatory and response variables.
2. *r* has no units and does not change when we change the units of measurement of *x*, *y*, or both.
3. Positive *r* indicates positive association between the variables, and negative *r* indicates negative association.
4. The correlation *r* is always a number between –1 and 1.

Cautions:

- Correlation requires that both variables be quantitative.
- Correlation *does not describe curved relationships* between variables, no matter how strong the relationship is.
- The correlation *r* is not resistant; it can be strongly affected by a few outlying observations.
- Correlation is not a complete summary of two-variable data.

Correlation Examples

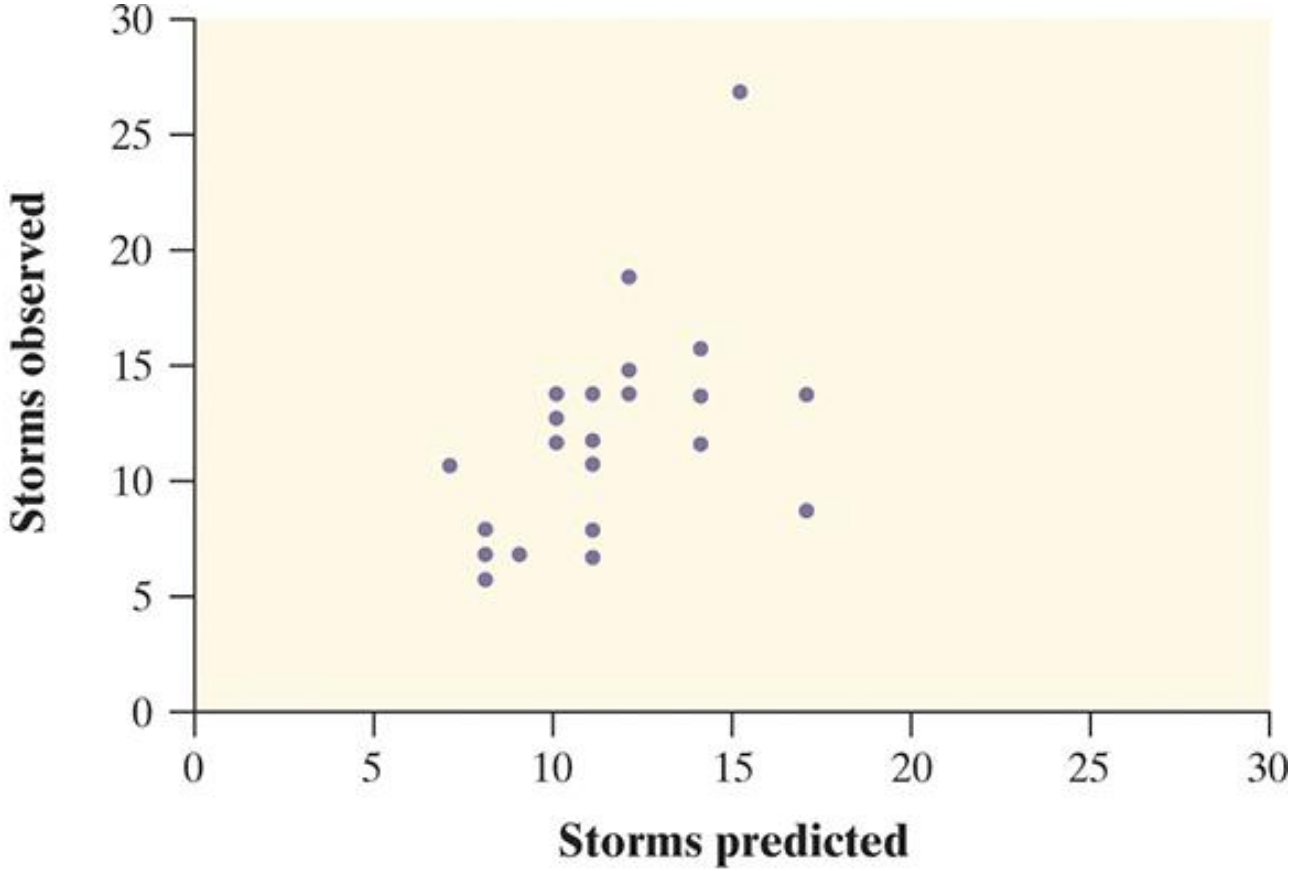
For each graph, estimate the correlation *r* and interpret it in context.



(a)

Correlation Examples

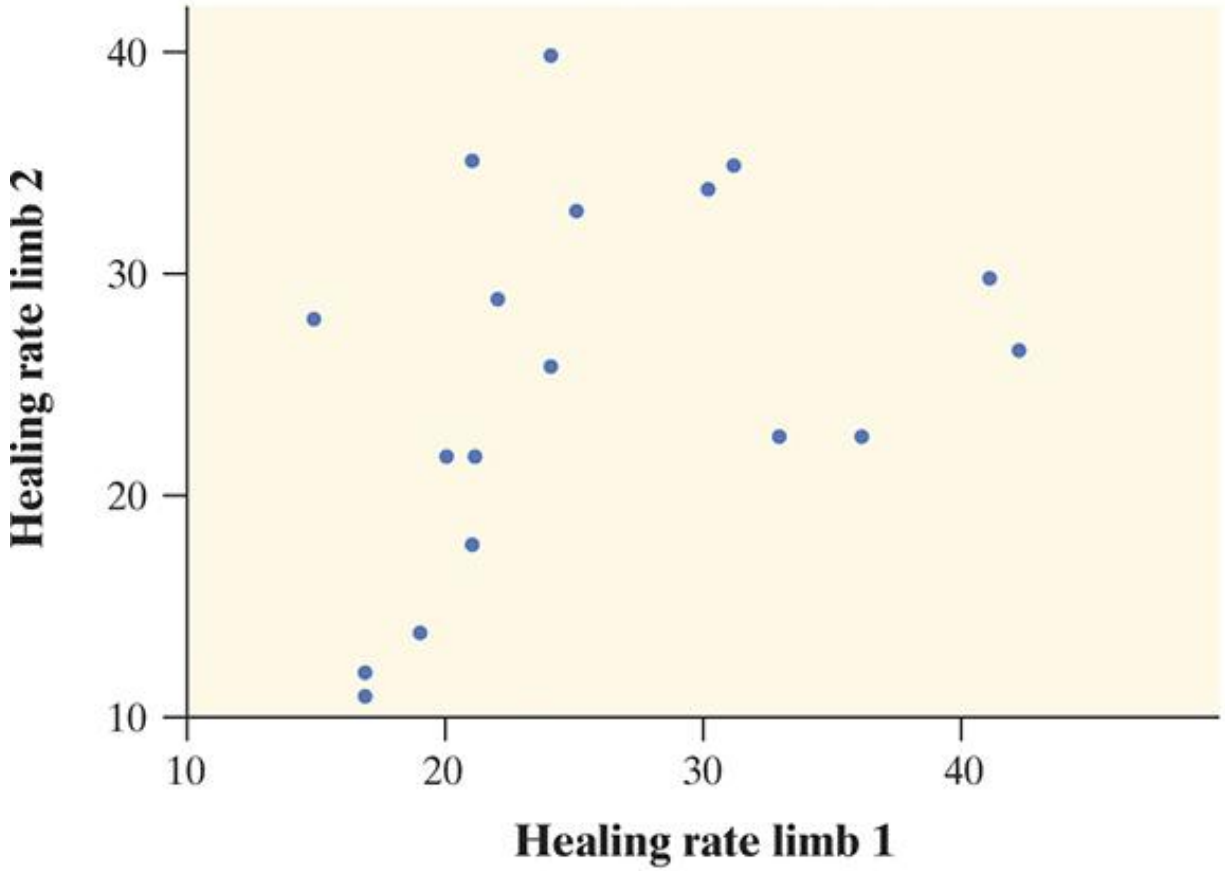
For each graph, estimate the correlation *r* and interpret it in context.



(b)

Correlation Examples

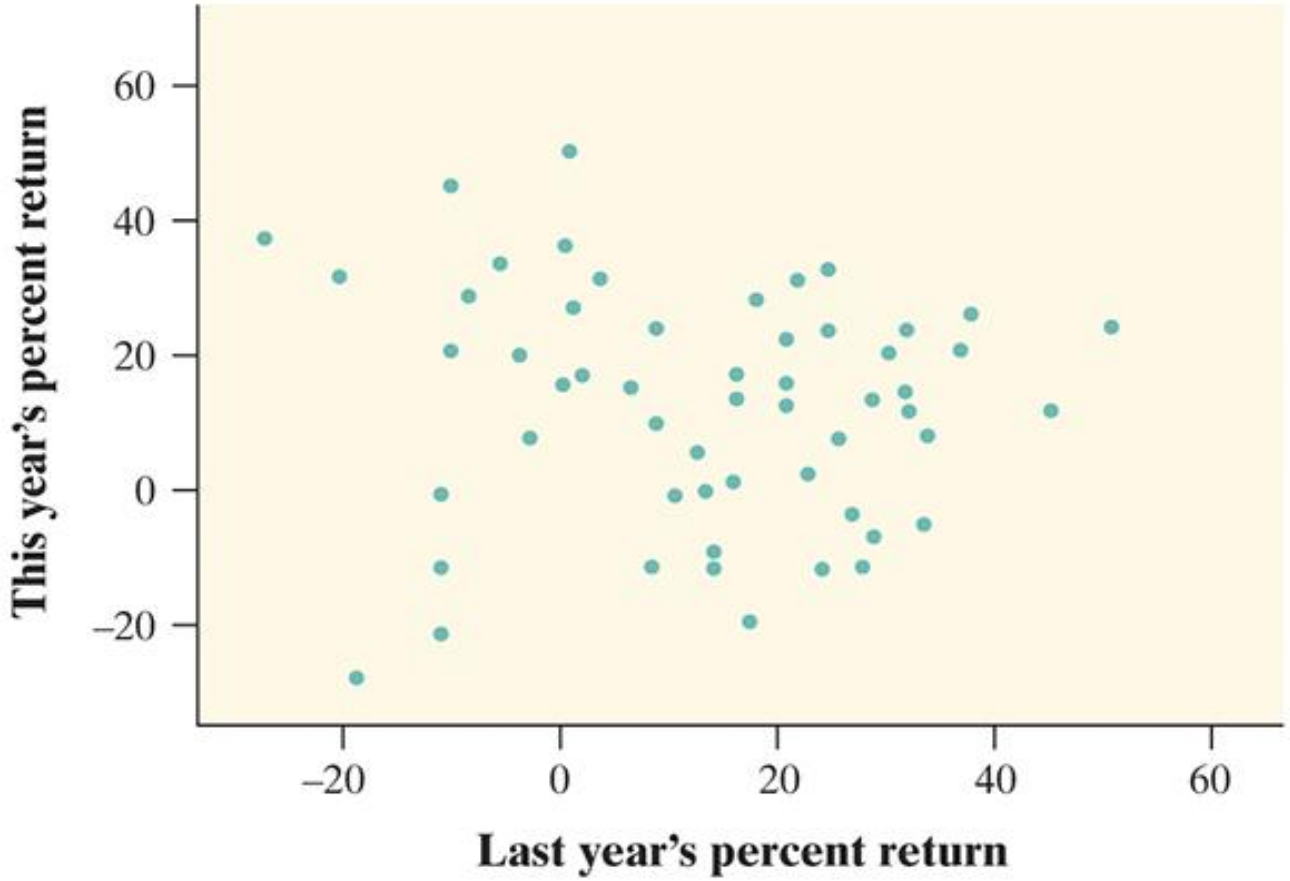
For each graph, estimate the correlation *r* and interpret it in context.



(c)

Correlation Examples

For each graph, estimate the correlation *r* and interpret it in context.



(d)

2.4 Least-Squares Regression

- Regression lines
- Least-squares regression line
- Facts about least-squares regression
- Correlation and regression

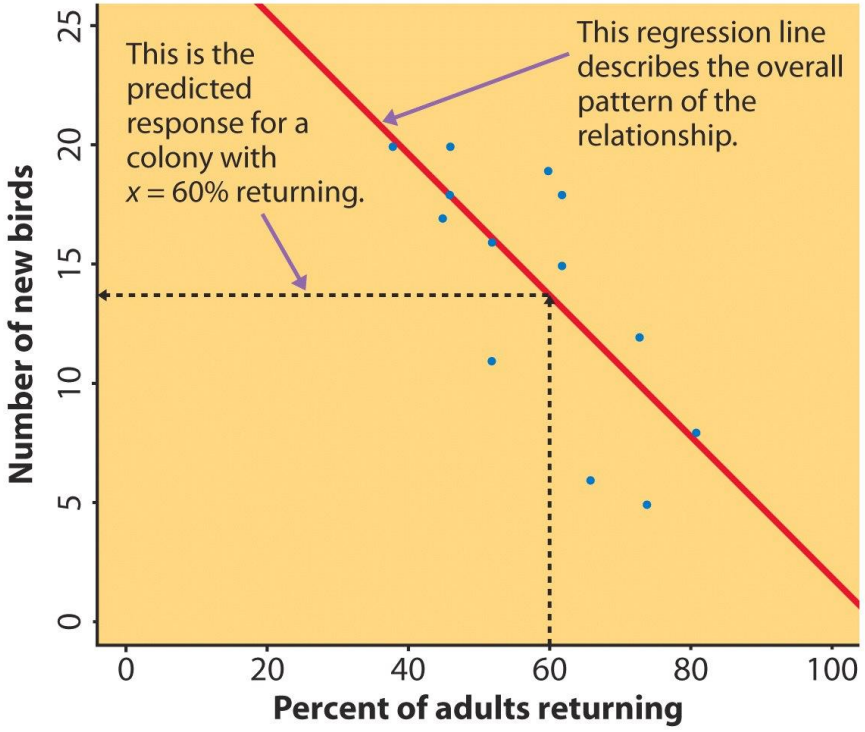
Regression Line

A **regression line** is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes.

We can use a regression line to predict the value of *y* for a given value of *x*.

Example: Predict the number of new adult birds that join the colony based on the percent of adult birds that return to the colony from the previous year.

- **If 60% of adults return, how many new birds are predicted?**



Regression Line

When a scatterplot displays a linear pattern, we can describe the overall pattern by drawing a straight line through the points. **Fitting a line** to data means drawing a line that comes as close as possible to the points.

Regression equation: $\hat{y} = b_0 + b_1 x$

- ***x*** is the value of the explanatory variable.
- **"*y*-hat"** is the predicted value of the response variable for a given value of *x*.
- b_1 is the **slope,** the amount by which *y* changes for each one-unit increase in *x.*
- b_0 is the **intercept,** the value of *y* when *x* = *0.*

Least-Squares Regression Line

Since we are trying to predict *y,* we want the regression line to be as close as possible to the data points in the vertical (*y*) direction.

Least-Squares Regression Line (LSRL):

The **least-squares regression line of *y* on *x*** is the line that minimizes the sum of the squares of the vertical distances of the data points from the line.

If we have data on an explanatory variable *x* and a response variable *y*, the equation of the least-squares regression line is:

$$y = b_0 + b_1 x$$

Facts About Least-Squares Regression

Regression is one of the most common statistical settings, and least-squares is the most common method for fitting a regression line to data. Here are some facts about least-squares regression lines.

- **Fact 1:** A change of one standard deviation in *x* corresponds to a change of *r* standard deviations in *y*.
- **Fact 2:** The LSRL always passes through (*x*-bar, *y*-bar)
- **Fact 3:** The distinction between explanatory and response variables is essential.

Correlation and Regression

Least-squares regression looks at the distances of the data points from the line only in the *y* direction. As a result, the variables *x* and *y* play different roles in regression. Even though correlation *r* ignores the distinction between *x* and *y,* there is a close connection between correlation and regression.

The **square of the correlation, r^2,** is the fraction of the variation in values of *y* that is explained by the least-squares regression of *y* on *x*.

- r^2 is called the **coefficient of determination.**

2.5 Cautions About Correlation and Regression

- Predictions
- Residuals and residual plots
- Outliers and influential observations
- Lurking variables
- Correlation and causation

Predictions Via Regression Line

For the returning birds example, the LSRL is:

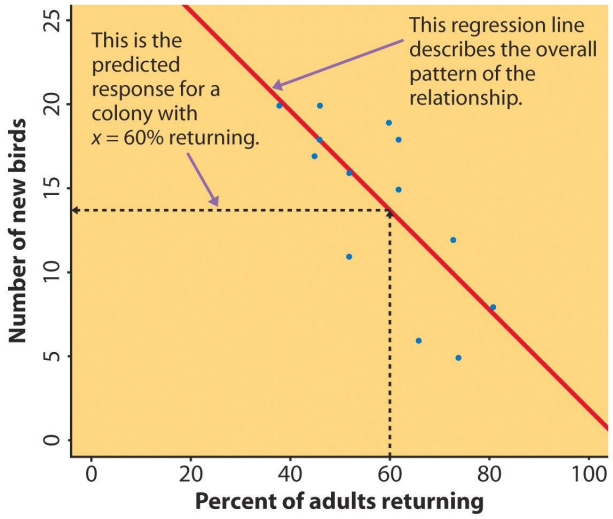
y*-hat = 31.9343 – 0.3040*x

y*-hat** is the predicted number of new birds for colonies with ***x percent of adults returning.

Suppose we know that an individual colony has 60% returning. What would we ***predict*** the number of new birds to be for just that colony?

For colonies with 60% returning, we ***predict*** the average number of new birds to be:

31.9343 – (0.3040)(60) = **13.69** birds



Residuals

A regression line describes the overall pattern of a linear relationship between an explanatory variable and a response variable. Deviations from the overall pattern are also important. The vertical distances between the points and the least-squares regression line are called *residuals*.

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line:

residual = observed *y* – predicted *y*

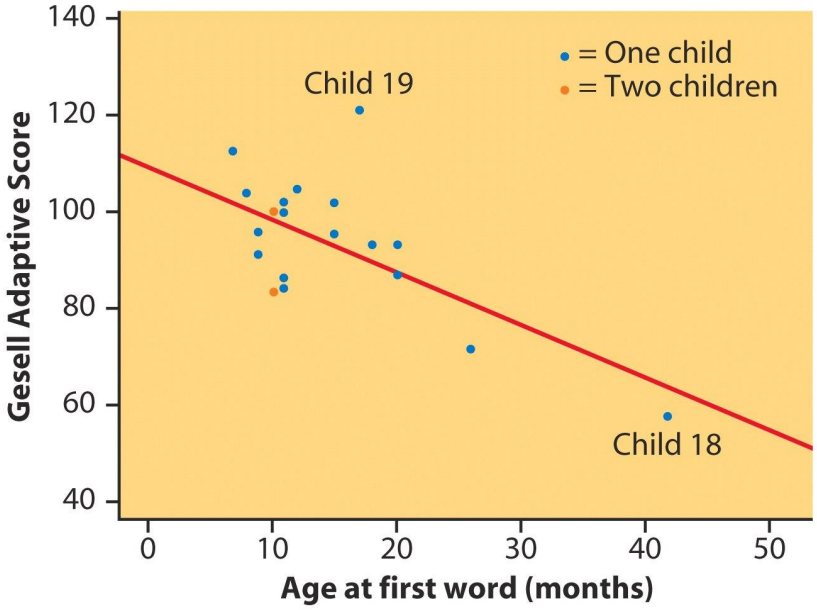
$$= y - \hat{y}$$

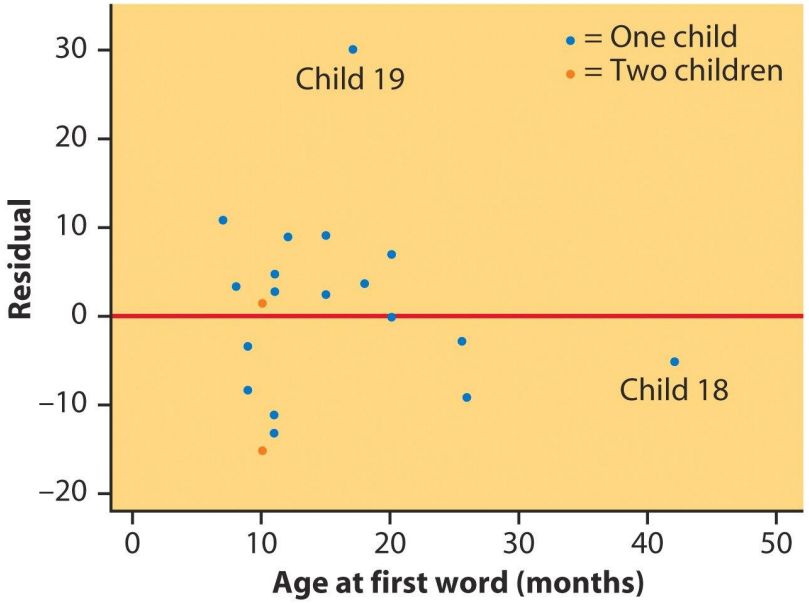
Residual Plots

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

- Ideally there should be a "random" scatter around zero.
- Residual *patterns* suggest deviations from a linear relationship.

Gesell Adaptive Score and Age at First Word



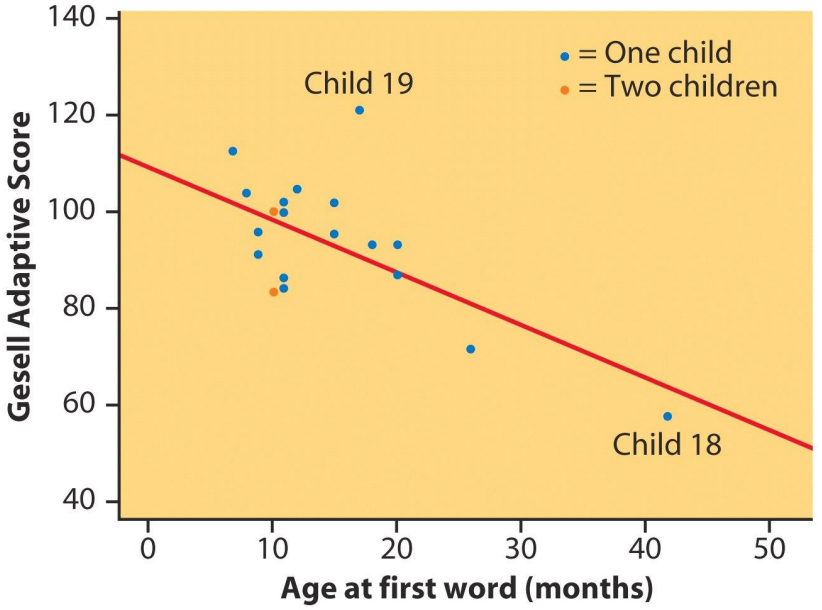


Outliers and Influential Points

An ***outlier*** is an observation that lies outside the overall pattern of the other observations.

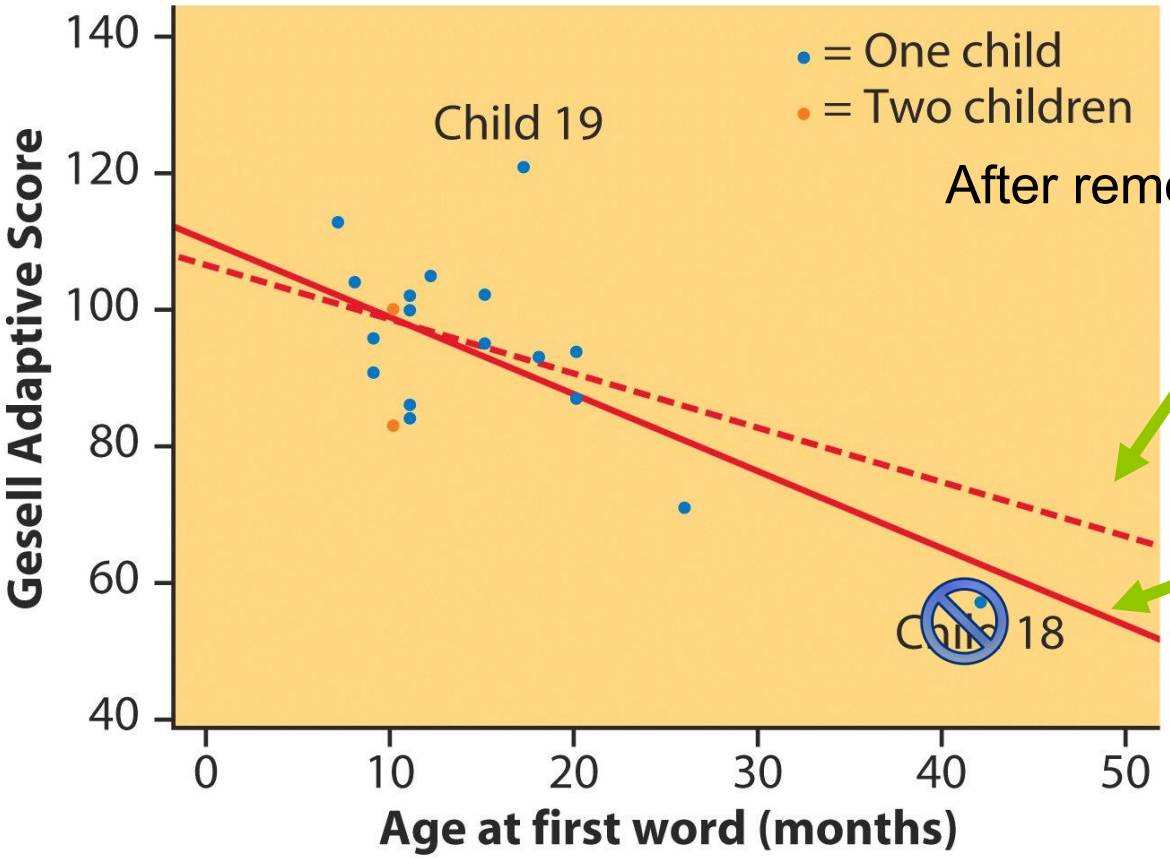
- Outliers in the *y* direction have large residuals.

Outliers in the *x* direction are often ***influential*** for the least-squares regression line, meaning that the removal of such points would markedly change the equation of the line.



Outliers and Influential Points

Gesell Adaptive Score and Age at First Word



After removing child 18

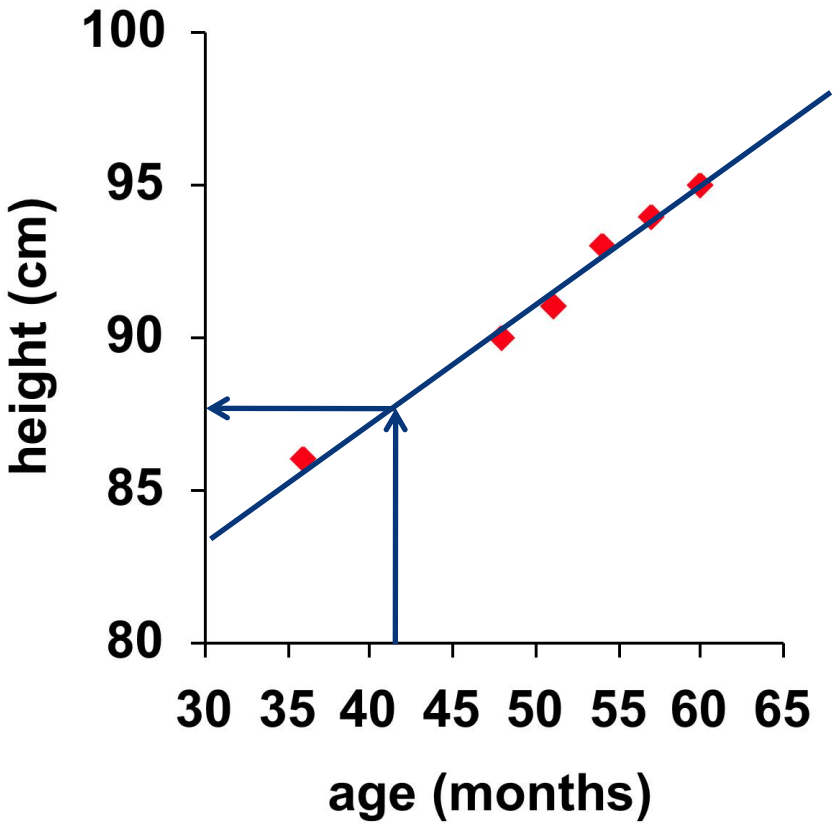
r^2 = 11%

From all of the data

r^2 = 41%

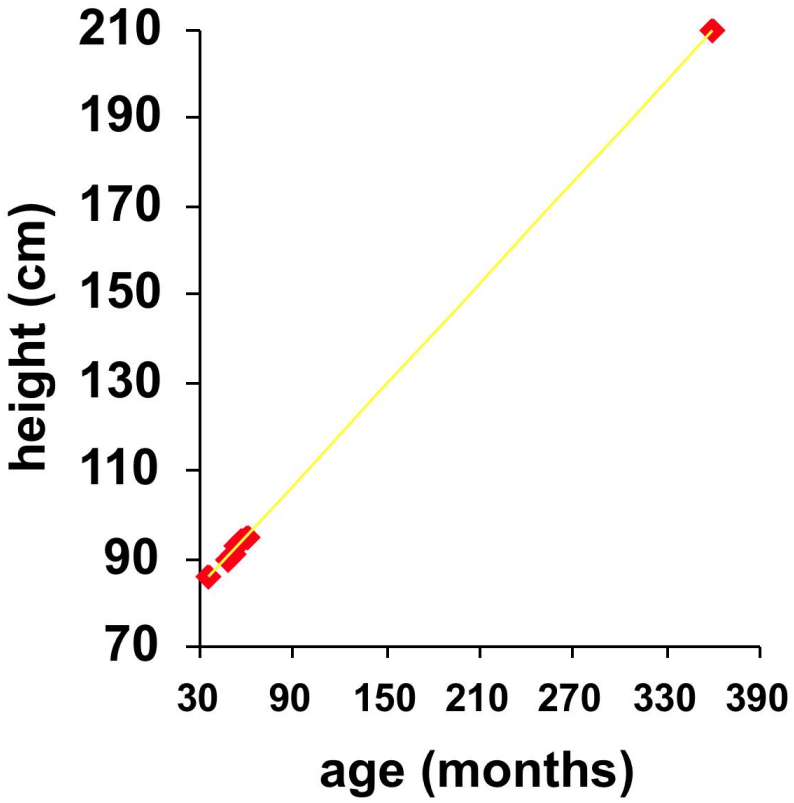
Extrapolation

- Sarah's height was plotted against her age.
- Can you guess (predict) her height at age 42 months?
- Can you predict her height at age 30 years (360 months)?



Extrapolation

- Regression line:
  y*-hat = 71.95 + .383 *x
- Height at age 42 months?
  ***y*-hat = 88**
- Height at age 30 years?
  ***y*-hat = 209.8**
- She is predicted to be 6’ 10.5” at age 30!
  What’s wrong?



Cautions About Correlation and Regression

- Both describe linear relationships.
- Both are affected by outliers.
- Always plot the data before interpreting.
- Beware of ***extrapolation.***
  - Use caution in predicting *y* when *x* is outside the range of observed *x*'s.
- Beware of ***lurking variables.***
  - These have an important effect on the relationship among the variables in a study, but are not included in the study.
- **Correlation does not imply causation!**

2.7 The Question of Causation

- Explaining association
- Causation
- Common response
- Confounding
- Establishing causation

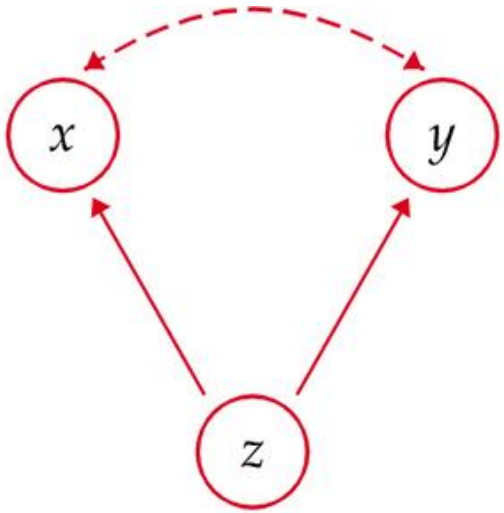
Explaining Association: Causation

Association, however strong, does NOT imply causation.

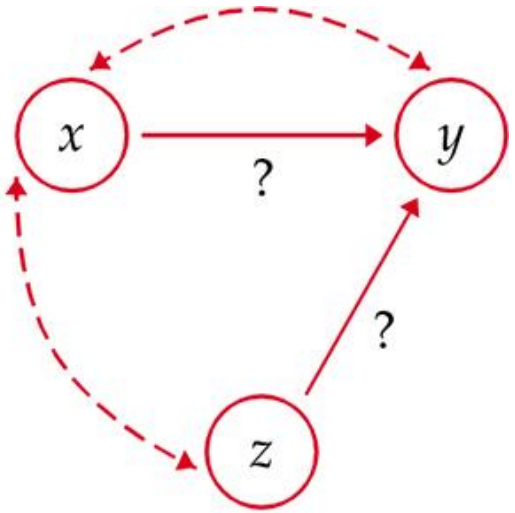
Some possible explanations for an observed association

The dashed lines show an association. The solid arrows show a cause-and-effect link. *x* is explanatory, *y* is response, and *z* is a lurking variable.

Causation



Common response



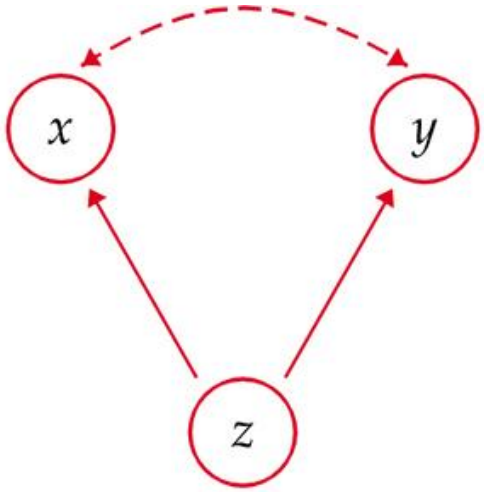
Confounding

Explaining Association: Common Response

"Beware the lurking variable" is good advice when thinking about an association between two variables. The observed relationship between the variables can be explained by a lurking variable. Both *x* and *y* may change in response to changes in *z*.

Most students who have high SAT scores (*x*) in high school have high GPAs (*y*) in their first year of college.

- This positive correlation can be explained as a *common response* to students' ability and knowledge.
- The observed association between *x* and *y* could be explained by a third lurking variable *z.* In this example, "ability and knowledge" is the lurking variable.
- Both *x* and *y* change in response to changes in *z*. This creates an association even though there is no direct causal link.

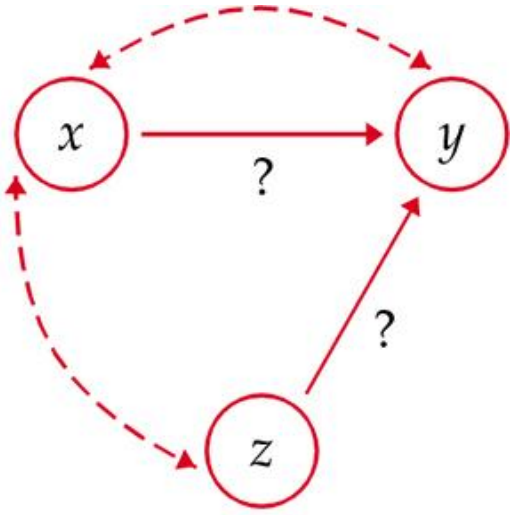


Common response

Explaining Association: Confounding

Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.

- Example: Studies have found that religious people live longer than nonreligious people.
- Religious people also take better care of themselves and are less likely to smoke or be overweight.



Confounding

Establishing Causation

It appears that lung cancer is associated with smoking.

How do we know that both of these variables are not being affected by an unobserved third (lurking) variable?

For instance, what if there is a genetic predisposition that causes people to both get lung cancer *and* become addicted to smoking, but the smoking itself doesn't CAUSE lung cancer?

We can evaluate the association using the following criteria:

1. The association is strong.
2. The association is consistent.
3. Higher doses are associated with stronger responses.
4. Alleged cause precedes the effect.
5. The alleged cause is plausible.

Evidence of Causation

A properly conducted ***experiment*** may establish causation.

Other considerations when we cannot do an experiment:

- The association is *strong.*
- The association is *consistent.*
  - The connection happens in *repeated trials.*
  - The connection happens under *varying conditions.*
- Higher doses are associated with stronger responses.
- Alleged cause *precedes* the effect in time.
- Alleged cause is *plausible* (reasonable explanation).

Chapter 2 Looking at Data—Relationships

2.1 Relationships

2.2 Scatterplots

2.3 Correlation

2.4 Least-Squares Regression

2.5 Cautions about Correlation and Regression

2.6 Data Analysis for Two-Way Tables

2.7 The Question of Causation