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

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.

2.2 Scatterplots

- Scatterplots
- Interpreting scatterplots
- Categorical variables in scatterplots

Scatterplot 1

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 cases. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each case 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.

Interpreting Scatterplots 1

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

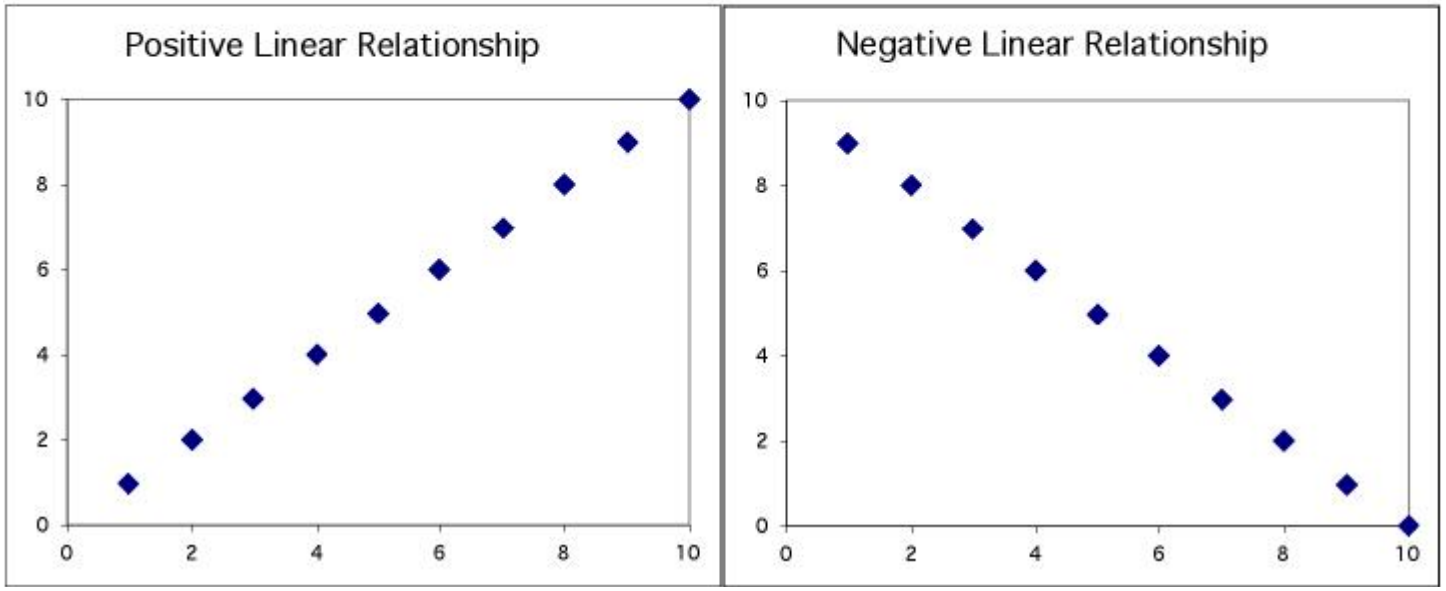
As in any graph of data, look for the *overall pattern* and for striking *deviations* from that pattern.

- You can describe the overall pattern of a scatterplot by the **form**, **direction**, 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.

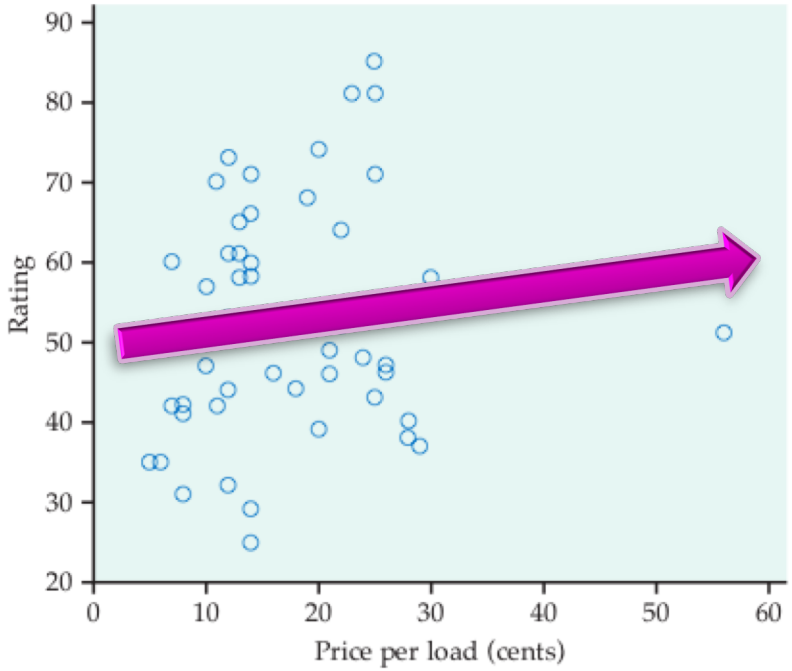
Interpreting Scatterplots 2

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 3



Outlier

- ✓ There is one possible outlier—the detergent with the price per load of 56 cents seems to have a lower rating than one might expect if higher price means a higher rated detergent.

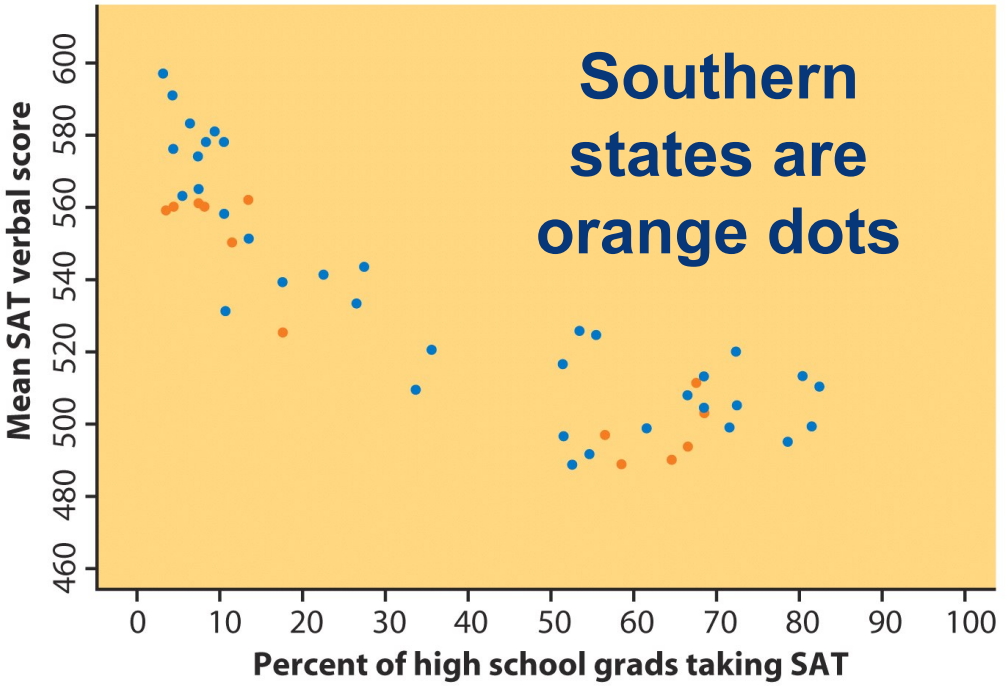
Strength **Direction** **Form**

- ✓ There is a moderately weak, positive, linear relationship between price per load and detergent rating.
- ✓ It appears that more expensive detergents are rated slightly higher, with possibly one exception.

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.



2.3 Correlation

- Correlation coefficient *r*
- Properties of *r*

Measuring Linear Association 1

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:

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

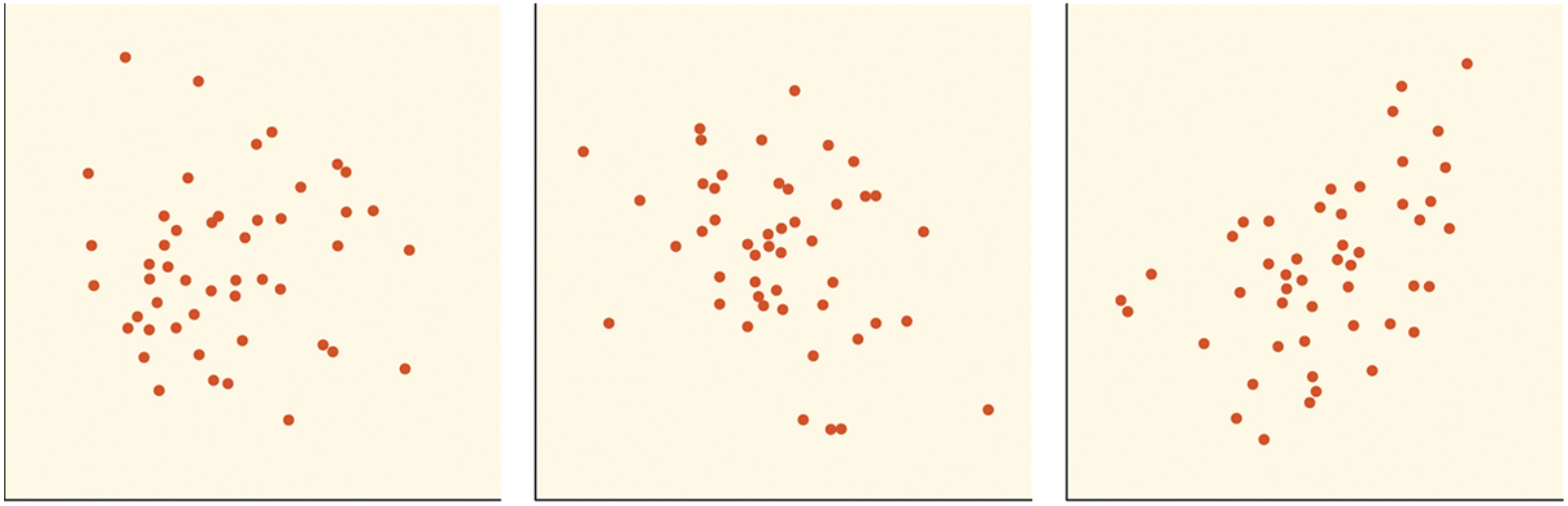
Measuring Linear Association 2

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.

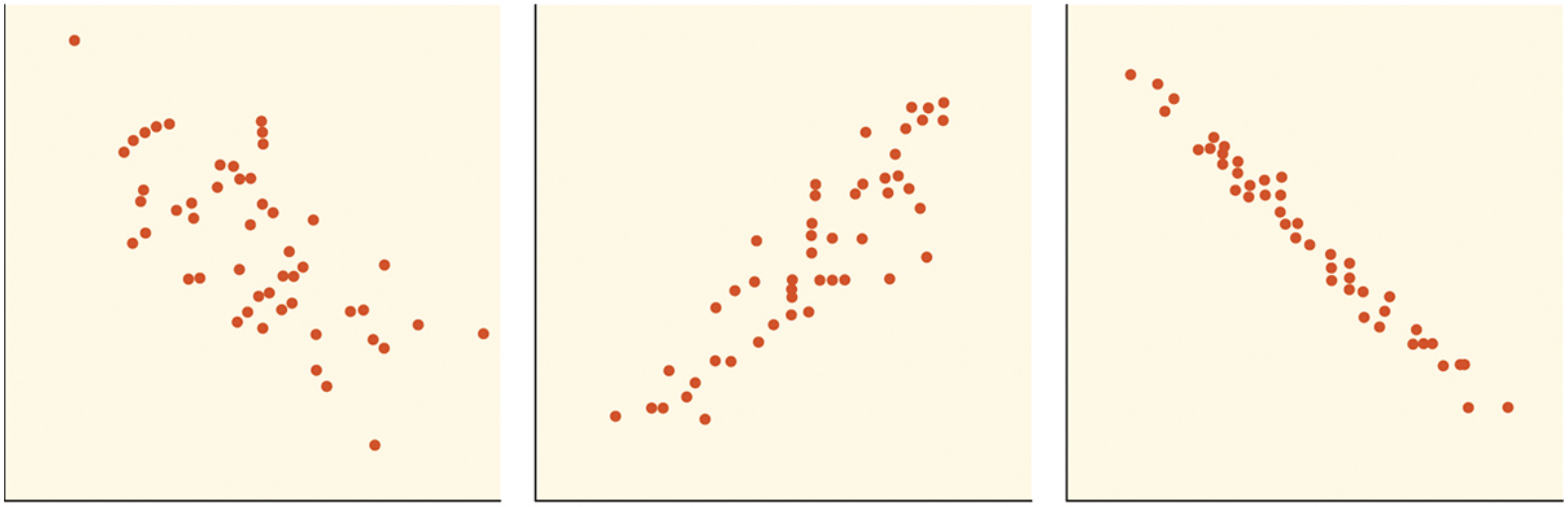
Properties of Correlation

- *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.

2.4 Least-Squares Regression

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

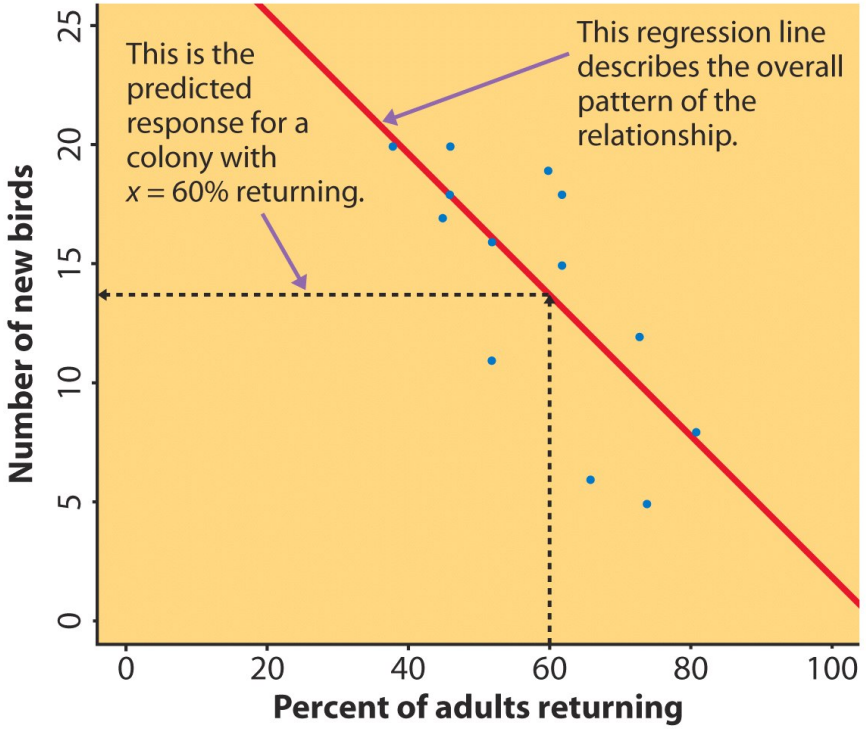
Regression Line 1

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 2

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

Because 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 LSRL is

$$\hat{y} = b_0 + b_1 x$$

Predictions Via Regression Line

For the returning birds example, the LSRL is

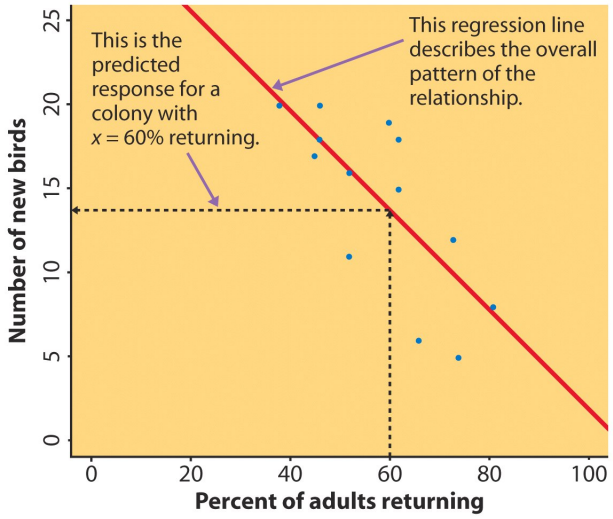
y*-hat = 31.9343 – 0.3040*x

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?

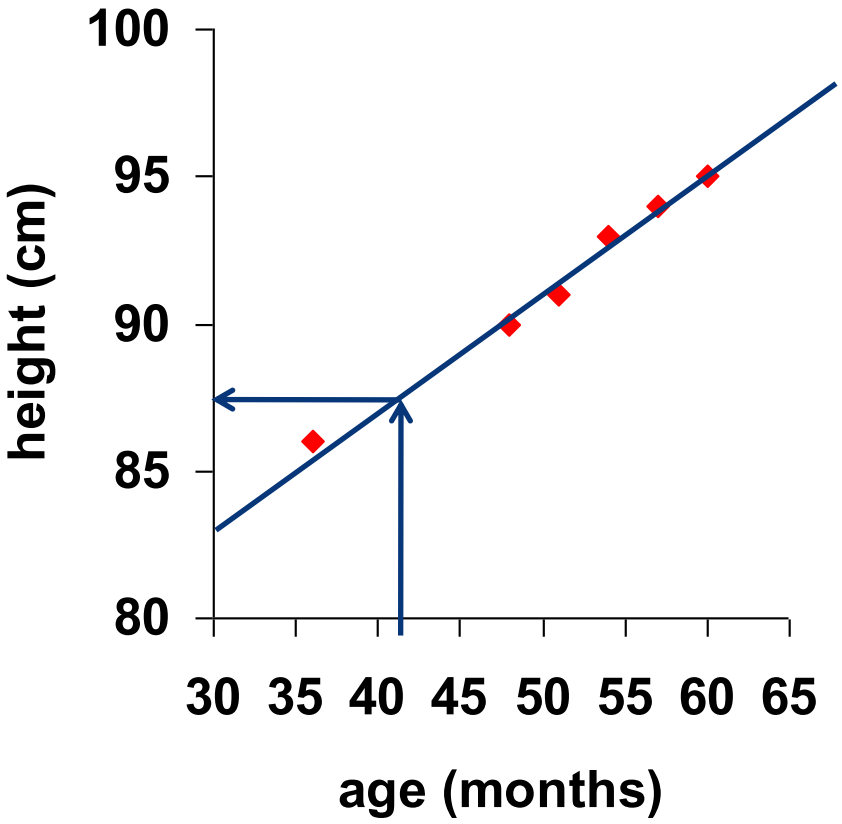
For colonies with 60% returning, we ***predict*** the average number of new birds to be

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



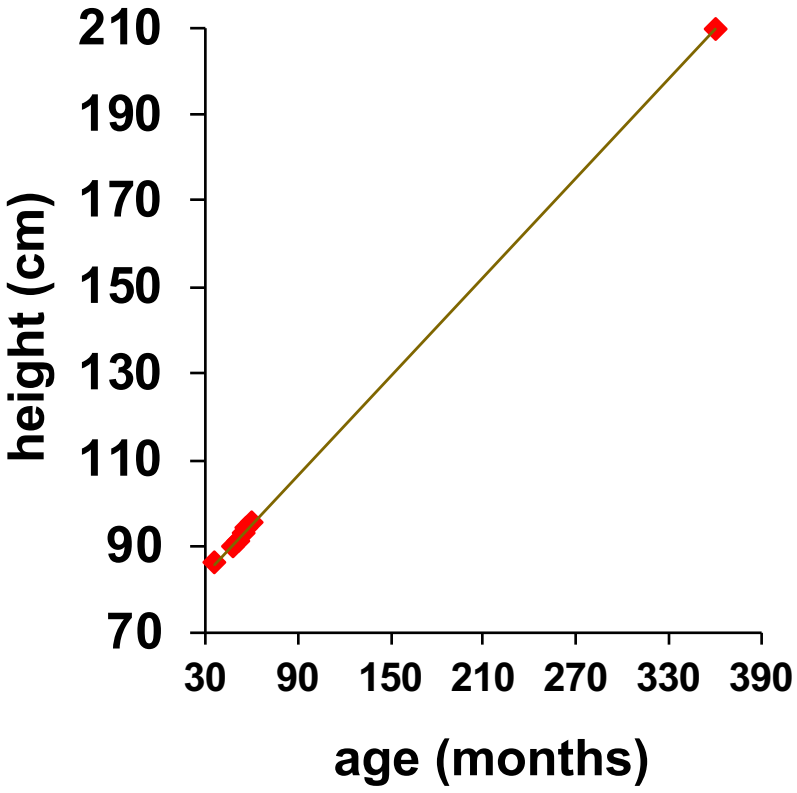
Extrapolation 1

- 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 2

- 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?*



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 sometimes called the **coefficient of determination.**

2.5 Cautions about Correlation and Regression

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

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*

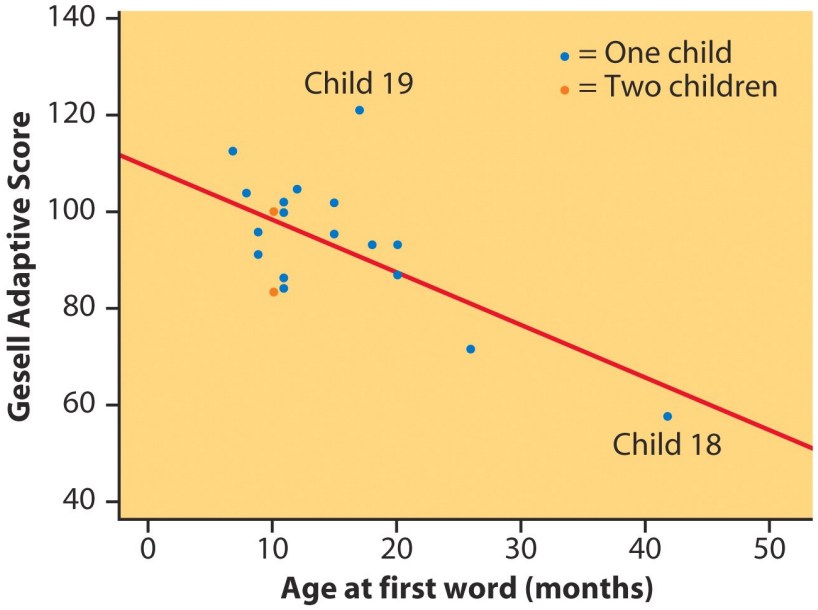
$$= y - \hat{y}$$

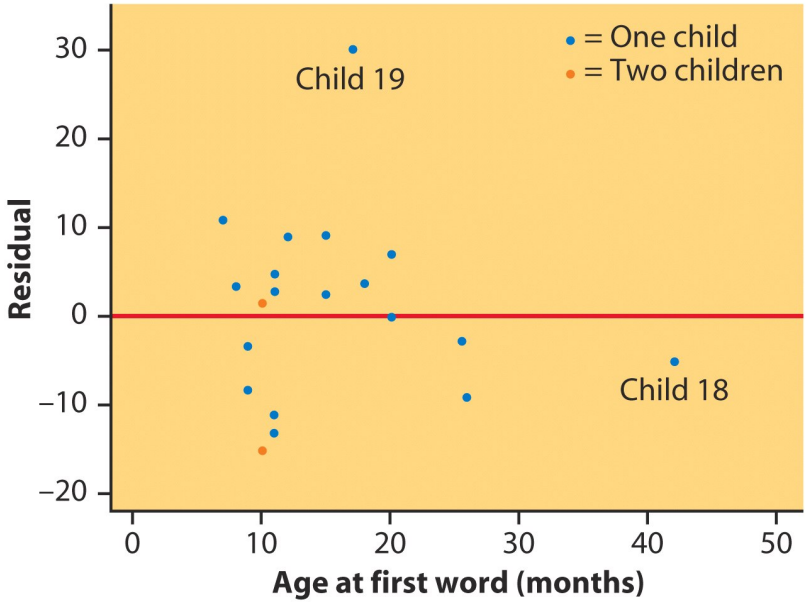
Residual Plots

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.

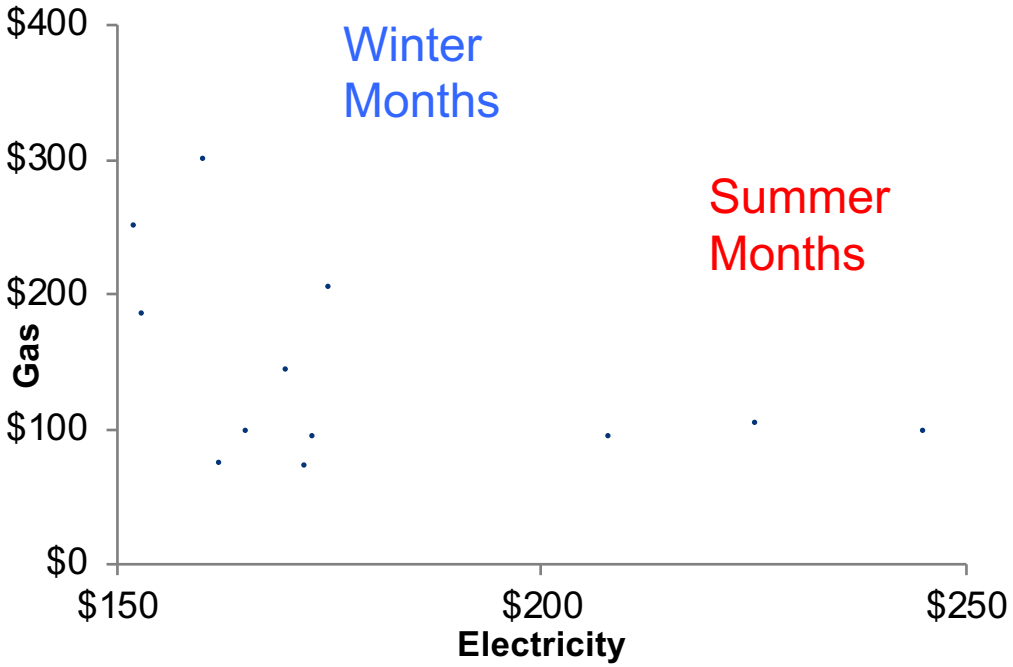
Gesell Adaptive Score and Age at First Word





Lurking Variables

A **lurking variable** is a variable that is not among the explanatory or response variables and yet may influence the interpretation of relationships among those variables.



Gas expense seems to be negatively associated with Electricity expense, but "season" is a lurking variable.

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!**

2.6 Data Analysis for Two-Way Tables

- The two-way table
- The Two-Way Table
- Marginal distributions
- Conditional distributions
- Simpson’s paradox

Categorical Variables

Recall that categorical variables place individuals into one of several groups or categories.

- The values of a categorical variable are labels for the different categories.
- The distribution of a categorical variable lists the count or percent of individuals who fall into each category.

When a dataset involves two categorical variables, we begin by examining the counts or percents in various categories for *one* of the variables.

A **two-way table** describes two categorical variables, organizing counts according to a **row variable** and a **column variable.** Each combination of values for these two variables is called a **cell.**

The Two-Way Table

Young adults by gender and chance of getting rich by age 30

	Female	Male	Total
Almost no chance	96	98	194
Some chance, but probably not	426	286	712
A 50–50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

What are the variables described by this two-way table?

How many young adults were surveyed?

Marginal Distribution 1

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among *all* individuals described by the table.

Note: Percentages are often more informative than counts, especially when comparing groups of different sizes.

To examine a marginal distribution:

1. Use the data in the table to calculate the marginal distribution (in percent) of the row or column totals.

2. Make a graph to display the marginal distribution.

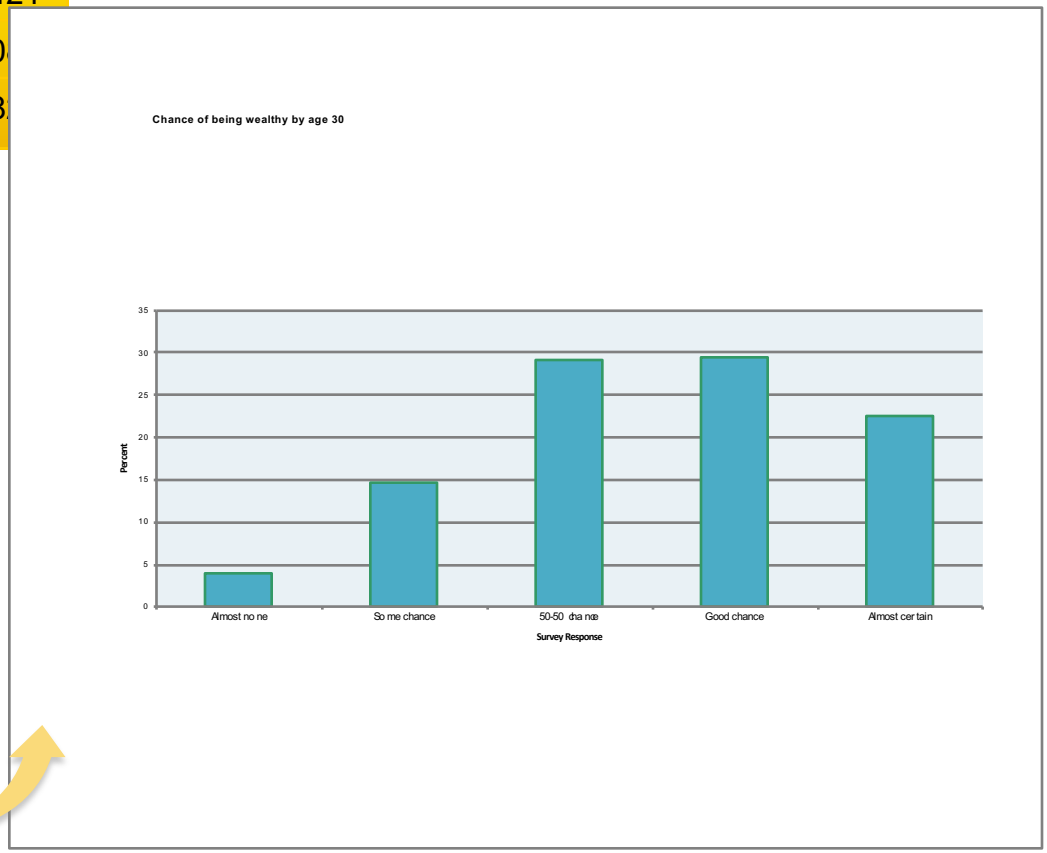
Marginal Distribution 2

Young adults by gender and chance of getting rich

	Female	Male	Total
Almost no chance	96	98	194
Some chance, but probably not	426	286	712
A 50–50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

Examine the **marginal distribution** of chance of getting rich.

Response	Percent
Almost no chance	194/4826 = 4.0%
Some chance	712/4826 = 14.8%
A 50–50 chance	1416/4826 = 29.3%
A good chance	1421/4826 = 29.4%
Almost certain	1083/4826 = 22.4%



Conditional Distribution 1

Marginal distributions tell us nothing about the relationship between two variables. For that, we need to explore the conditional distributions of the variables.

A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable.

To examine or compare conditional distributions:

1. Select the row(s) or column(s) of interest.
2. Use the data in the table to calculate the conditional distribution (in percent) of the row(s) or column(s).
3. Make a graph to display the conditional distribution.
   - Use a **side-by-side bar graph** or **segmented bar graph** to compare distributions.

Conditional Distribution 2

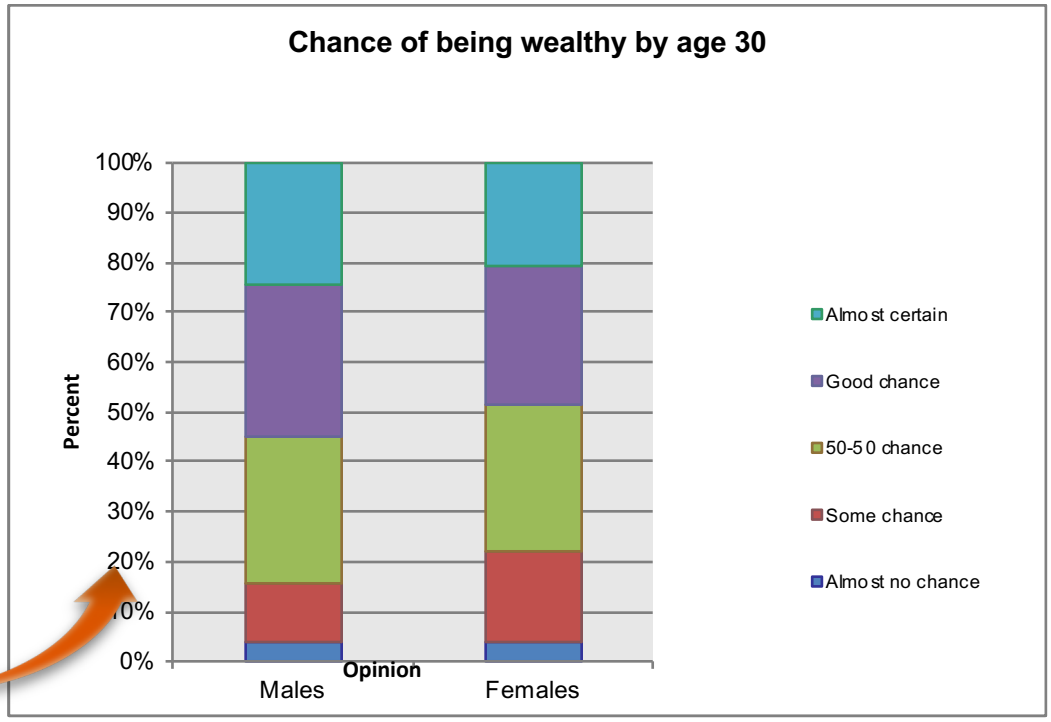
Young adults by gender and [illegible] getting rich

	Female	Male	Total
Almost no chance	96	98	194
Some chance, but probably not	426	286	712
A 50–50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

Calculate the conditional distribution of opinion among males.

Examine the relationship between gender and opinion.

Response	Male	Female
Almost no chance	98/2459 = 4.0%	96/2367 = 4.1%
Some chance	286/2459 = 11.6%	426/2367 = 18.0%
A 50–50 chance	720/2459 = 29.3%	696/2367 = 29.4%
A good chance	758/2459 = 30.8%	663/2367 = 28.0%
Almost certain	597/2459 = 24.3%	[illegible]/2367 = 20.5%



Simpson's Paradox 1

When studying the relationship between two variables, there may exist a **lurking variable** that creates a reversal in the direction of the relationship when the lurking variable is ignored as opposed to the direction of the relationship when the lurking variable is considered.

The lurking variable creates subgroups, and failure to take these subgroups into consideration can lead to misleading conclusions regarding the association between the two variables.

An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox.**

Simpson's Paradox 2

Consider the acceptance rates for the following groups of men and women who applied to college.

Counts	Accepted	Not accepted	Total
Men	198	162	360
Women	88	112	200
Total	286	274	560

Percents	Accepted	Not accepted
Men	**55%**	45%
Women	44%	56%

A higher percentage of **men** were accepted: Is there evidence of discrimination?

Simpson's Paradox 3

Consider the acceptance rates when broken down by type of school.

BUSINESS SCHOOL

Counts	Accepted	Not accepted	Total
Men	18	102	120
Women	24	96	120
Total	42	198	240

Percents	Accepted	Not accepted
Men	15%	85%
Women	**20%**	80%

ART SCHOOL

Counts	Accepted	Not accepted	Total
Men	180	60	240
Women	64	16	80
Total	244	76	320

Percents	Accepted	Not accepted
Men	75%	25%
Women	80%	20%

Simpson's Paradox 4

- **Lurking variable:** Applications were split between the Business School (240) and the Art School (320).

  Within each school a higher percentage of women were accepted than men.

  There is not any discrimination against women!

  This is an example of **Simpson's paradox.**

  When the lurking variable (Type of School: Business or Art) is ignored, the data seem to suggest discrimination against women.

  However, when the type of school is considered, the association is reversed and suggests discrimination against men.