Chapter 3: Describing Relationships

The Practice of Statistics, 4th edition – For AP*

STARNES, YATES, MOORE

+

Chapter 3
Describing Relationships

- **3.1 Scatterplots and Correlation**
- **3.2** Least-Squares Regression

Learning Targets

After this section, you should be able to…

- ✓ IDENTIFY explanatory and response variables
- ✓ CONSTRUCT scatterplots to display relationships
- ✓ INTERPRET scatterplots
- ✓ MEASURE linear association using correlation
- ✓ INTERPRET correlation

Explanatory and Response Variables

Most statistical studies examine data on more than one variable. In many of these settings, the two variables play different roles.

Definition:

A **response variable** measures an outcome of a study. An **explanatory variable** may help explain or influence changes in a response variable.

Note: In many studies, the goal is to show that changes in one or more explanatory variables actually *cause* changes in a response variable. However, other explanatory-response relationships don't involve direct causation.

Displaying Relationships: Scatterplots

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

Definition:

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 in the data appears as a point on the graph.

How to Make a Scatterplot

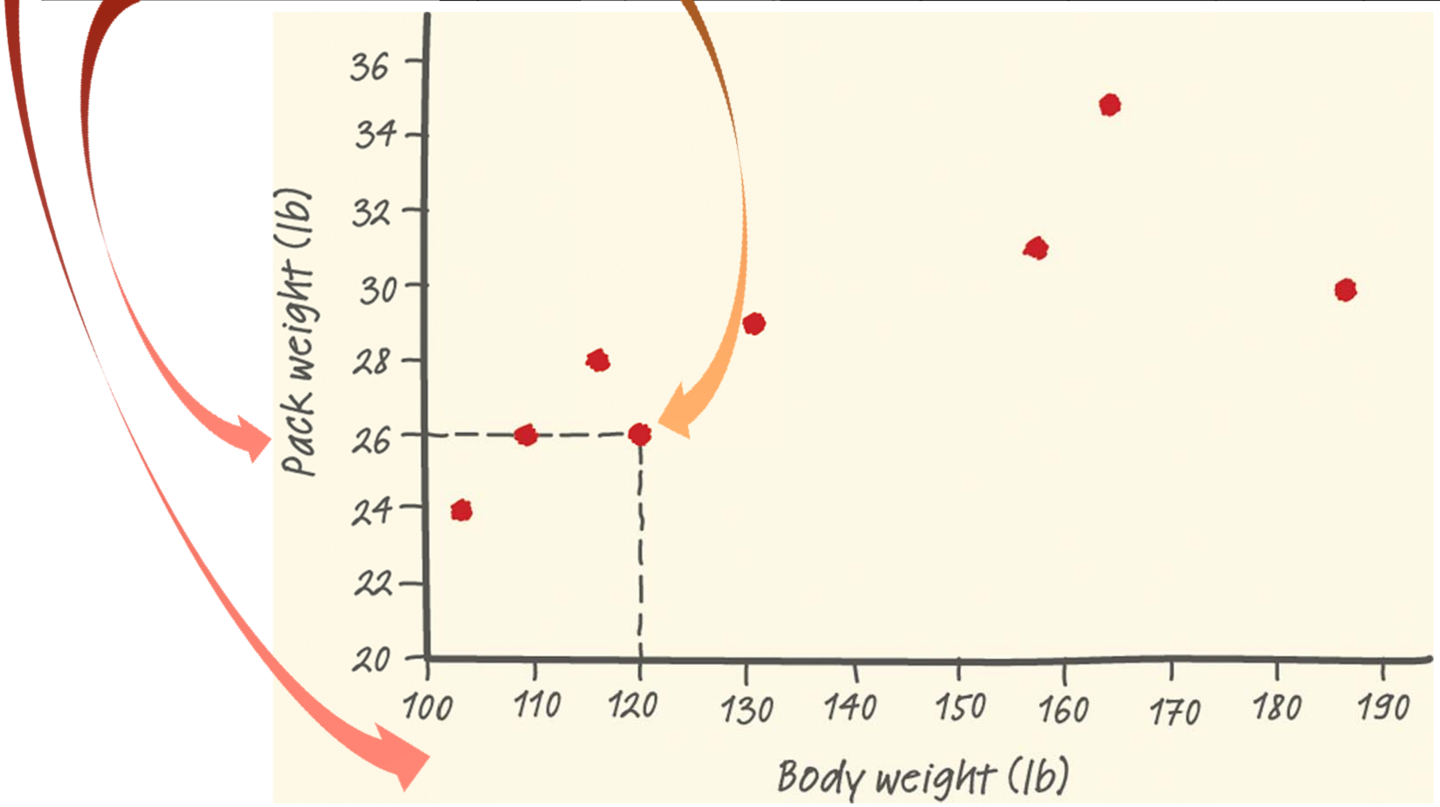
1. Decide which variable should go on each axis.
   - *Remember, the eXplanatory variable goes on the X-axis!*
2. Label and scale your axes.
3. Plot individual data values.

Displaying Relationships: Scatterplots

Make a scatterplot of the relationship between body weight and pack weight.

Since Body weight is our eXplanatory variable, be sure to place it on the X-axis!

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



Interpreting Scatterplots

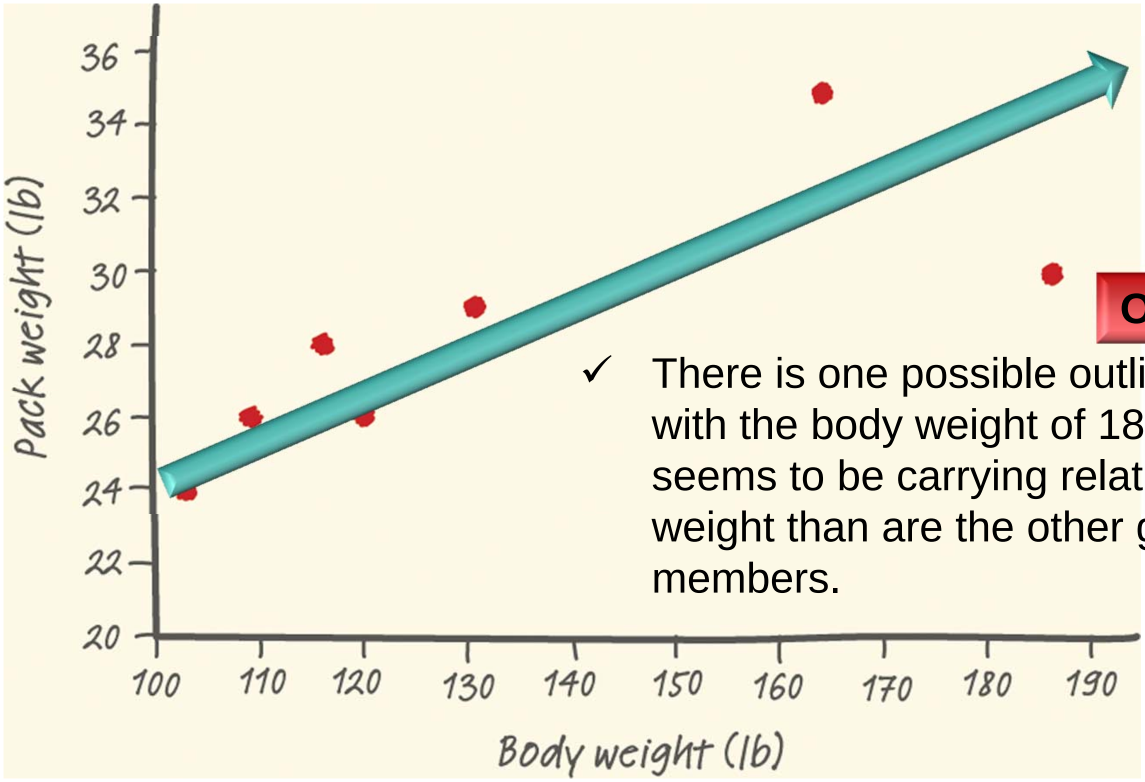
To interpret a scatterplot, follow the basic strategy of data analysis from Chapters 1 and 2. Look for patterns and important departures from those patterns.

How to Examine a Scatterplot

As in any graph of data, look for the *overall pattern* and for striking *departures* 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



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**

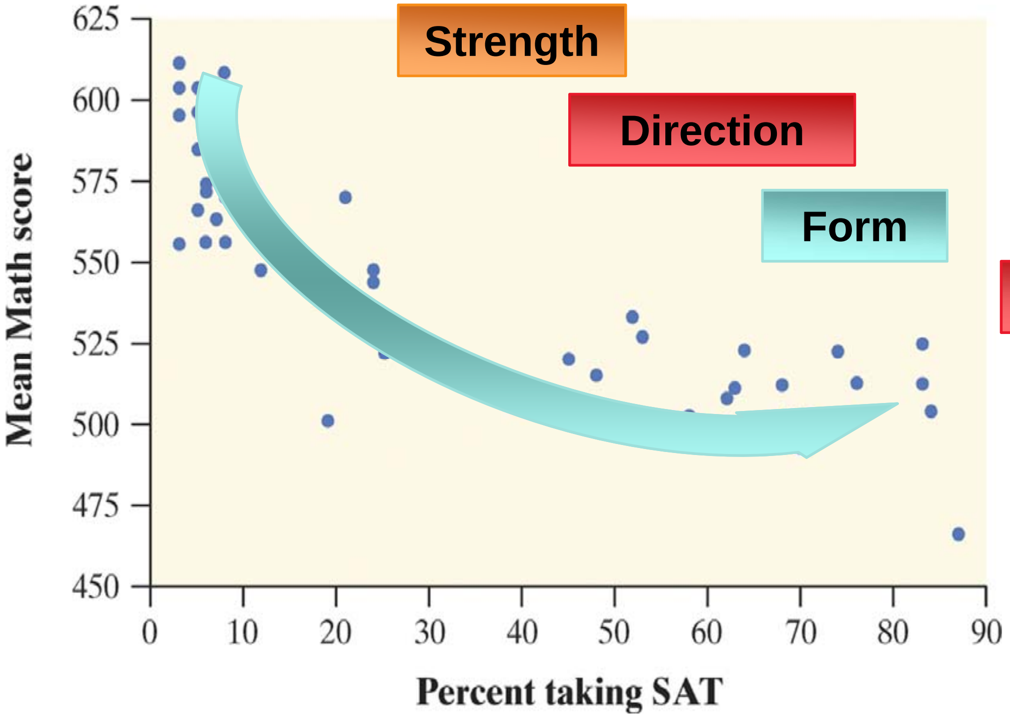
- ✓ There is a moderately strong, positive, linear relationship between body weight and pack weight.
- ✓ It appears that lighter students are carrying lighter backpacks.

Interpreting Scatterplots

Definition:

Two variables have a **positive association** 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 have a **negative association** when above-average values of one tend to accompany below-average values of the other.



Strength

Direction

Form

Consider the SAT example from page 144. Interpret the scatterplot.

There is a moderately strong, negative, curved relationship between the percent of students in a state who take the SAT and the mean SAT math score.
Further, there are two distinct clusters of states and two possible outliers that fall outside the overall pattern.

Measuring Linear Association: Correlation

A scatterplot displays the strength, direction, and form of the relationship between two quantitative variables.

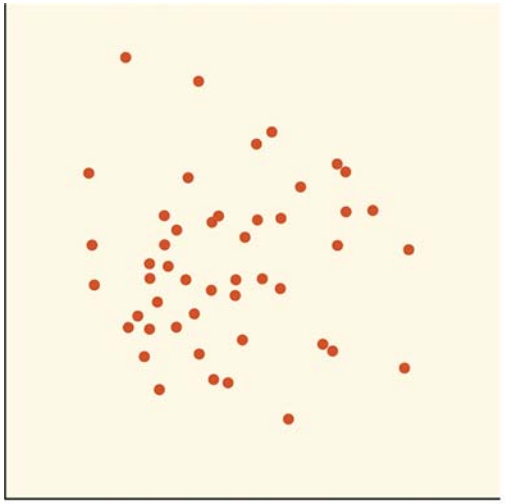
Linear relationships are important because a straight line is a simple pattern that is quite common. Unfortunately, our eyes are not good judges of how strong a linear relationship is.

Definition:

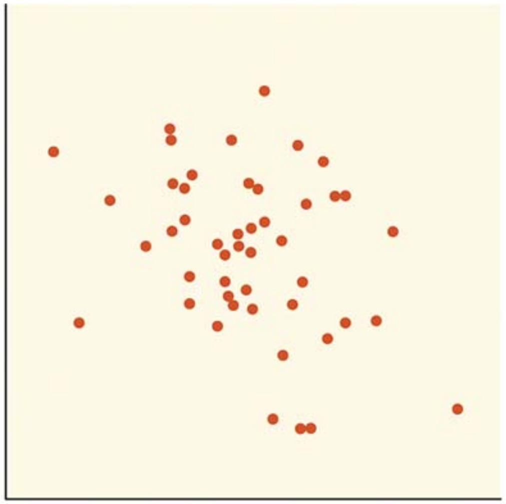
The **correlation *r*** measures the strength of the linear relationship between two quantitative variables.

- *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 towards -1 or 1.
- The extreme values *r* = -1 and r = 1 occur only in the case of a perfect linear relationship.

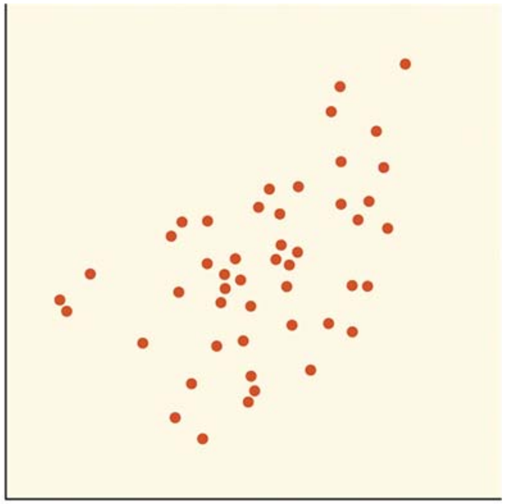
Measuring Linear Association: Correlation



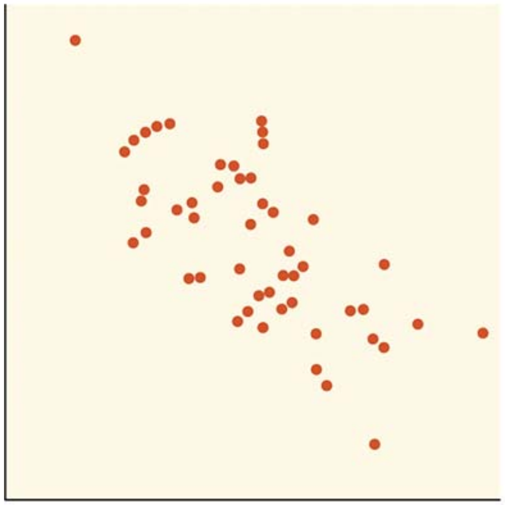
Correlation $r = 0$



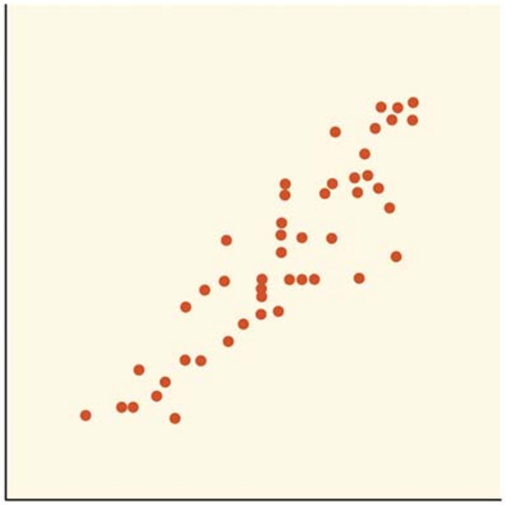
Correlation $r = -0.3$



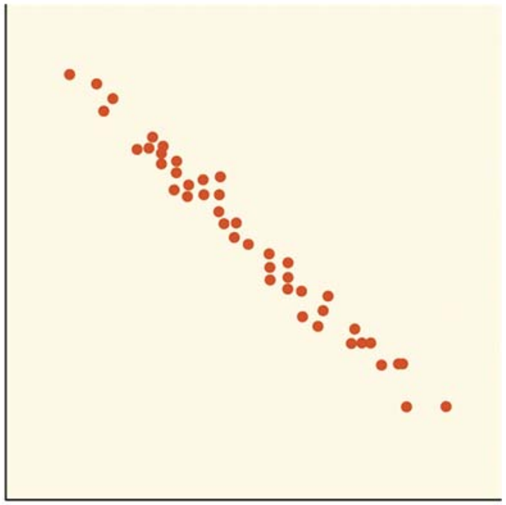
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

Correlation

The formula for *r* is a bit complex. It helps us to see what correlation is, but in practice, you should use your calculator or software to find *r*.

How to Calculate the Correlation *r*

Suppose that we have data on variables *x* and *y* for *n* individuals.

The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on.

The means and standard deviations of the two variables are *x-bar* and s_x for the *x*-values and *y-bar* and s_y for the *y*-values.

The correlation *r* between *x* and *y* is:

$$r = \frac{1}{n-1}\left[\left(\frac{x_1-\bar{x}}{s_x}\right)\left(\frac{y_1-\bar{y}}{s_y}\right)+\left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right)+\ldots+\left(\frac{x_n-\bar{x}}{s_x}\right)\left(\frac{y_n-\bar{y}}{s_y}\right)\right]$$

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

Facts about Correlation

How correlation behaves is more important than the details of the formula. Here are some important facts about *r*.

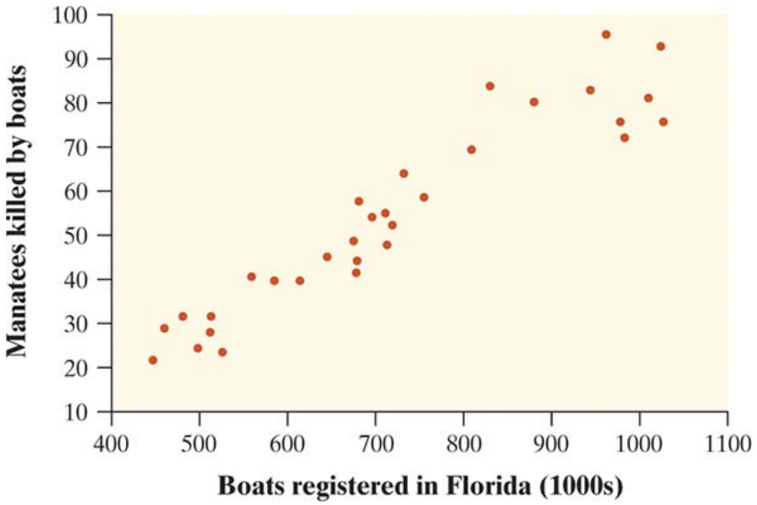
1. **Correlation makes no distinction between explanatory and response variables.**
2. ***r* does not change when we change the units of measurement of *x*, *y*, or both.**
3. **The correlation *r* itself has no unit of measurement.**

Cautions:

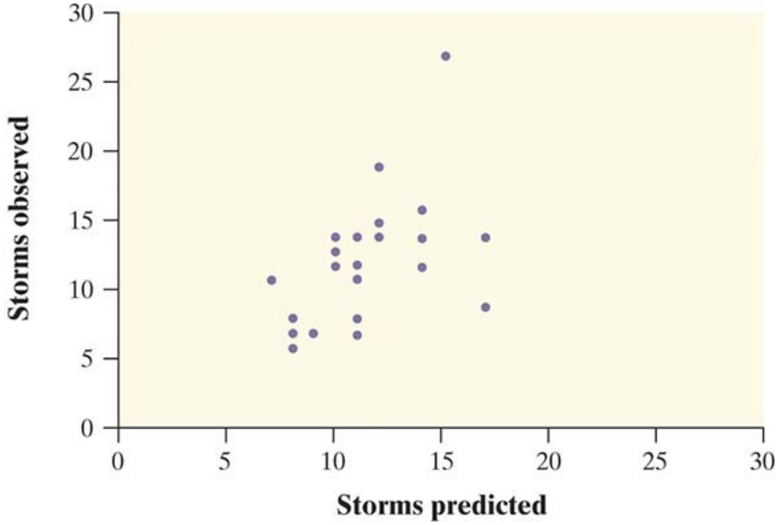
- Correlation requires that both variables be quantitative.
- Correlation does not describe curved relationships between variables, no matter how strong the relationship is.
- Correlation is not resistant. *r* is strongly affected by a few outlying observations.
- Correlation is not a complete summary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

Correlation Practice

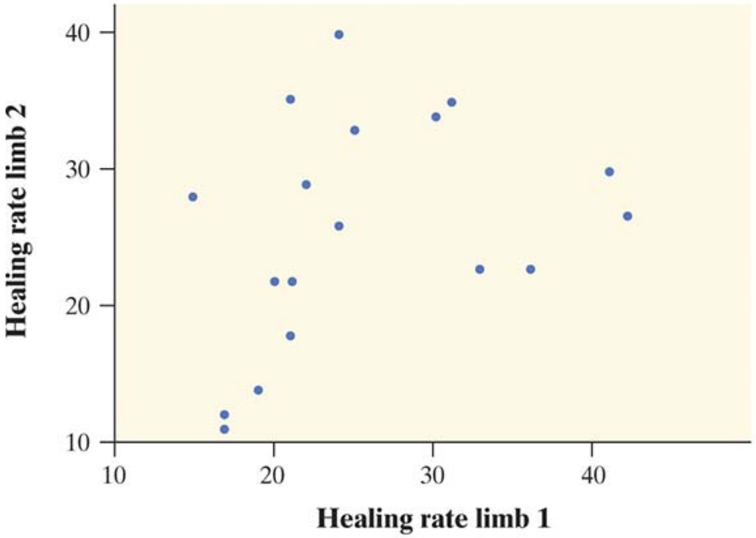
For each graph, estimate the correlation *r* and interpret it in context.



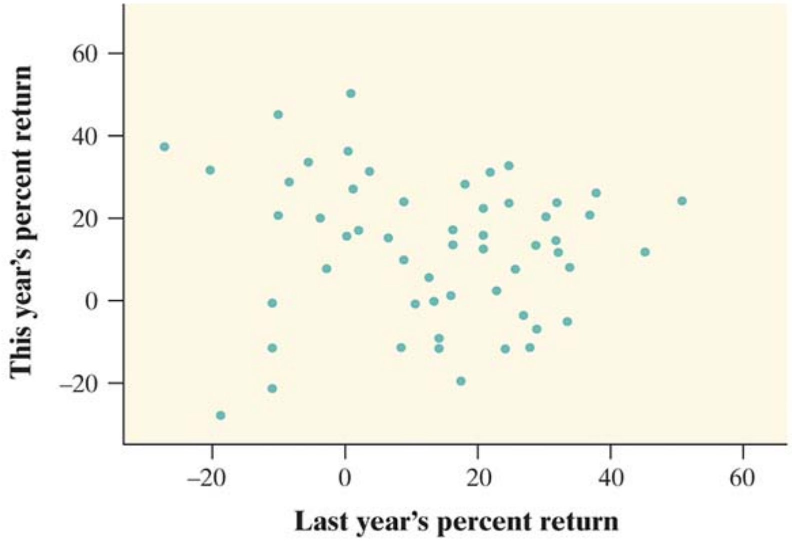
(a)



(b)



(c)



(d)

+

Section 3.1
Scatterplots and Correlation

Summary

In this section, we learned that…

- ✓ A **scatterplot** displays the relationship between two quantitative variables.
- ✓ An **explanatory variable** may help explain, predict, or cause changes in a **response variable.**
- ✓ When examining a scatterplot, look for an overall pattern showing the **direction**, **form**, and **strength** of the relationship and then look for **outliers** or other departures from the pattern.
- ✓ The **correlation *r*** measures the strength and direction of the linear relationship between two quantitative variables.

Section 3.2
Least-Squares Regression

Learning Objectives

After this section, you should be able to…

- ✓ INTERPRET a regression line
- ✓ CALCULATE the equation of the least-squares regression line
- ✓ CALCULATE residuals
- ✓ CONSTRUCT and INTERPRET residual plots
- ✓ DETERMINE how well a line fits observed data
- ✓ INTERPRET computer regression output

Least-Squares Regression

Regression Line

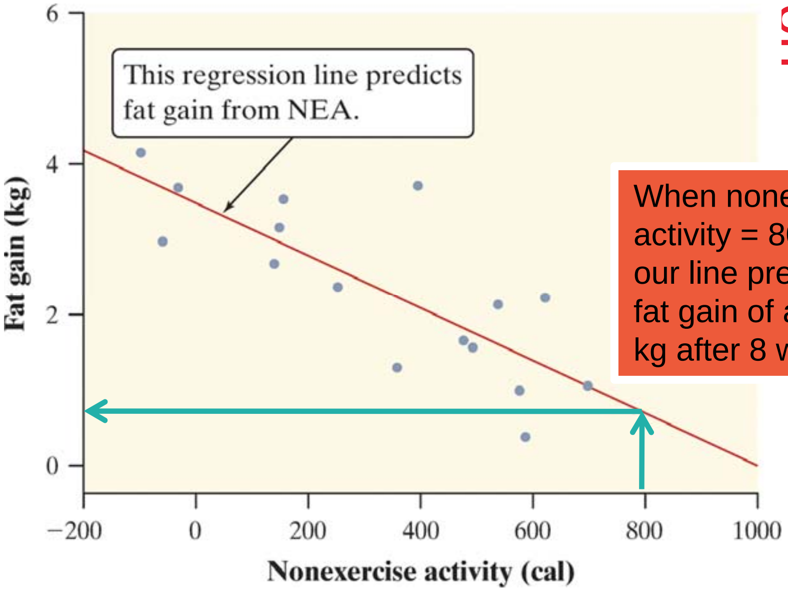
Linear (straight-line) relationships between two quantitative variables are common and easy to understand. A **regression line** summarizes the relationship between two variables, but only in settings where one of the variables helps explain or predict the other.

Definition:

A **regression line** is a line that describes how a response variable *y* changes as an explanatory variable *x* changes. We often use a regression line to predict the value of *y* for a given value of *x*.

Figure 3.7 on page 165 is a scatterplot of the change in nonexercise activity (cal) and measured fat gain (kg) after 8 weeks for 16 healthy young adults.

- ✓ The plot shows a moderately strong, negative, linear association between NEA change and fat gain with no outliers.
- ✓ The regression line predicts fat gain from change in NEA.



When nonexercise activity = 800 cal, our line predicts a fat gain of about 0.8 kg after 8 weeks.

Least-Squares Regression

Interpreting a Regression Line

A regression line is a *model* for the data, much like density curves. The equation of a regression line gives a compact mathematical description of what this model tells us about the relationship between the response variable *y* and the explanatory variable *x.*

Definition:

Suppose that *y* is a response variable (plotted on the vertical axis) and *x* is an explanatory variable (plotted on the horizontal axis). A **regression line** relating *y* to *x* has an equation of the form

$$\hat{y} = a + bx$$

In this equation,

•$\hat{y}$ (read "y hat") is the **predicted value** of the response variable *y* for a given value of the explanatory variable x.

•*b* is the **slope**, the amount by which *y* is predicted to change when *x* increases by one unit.

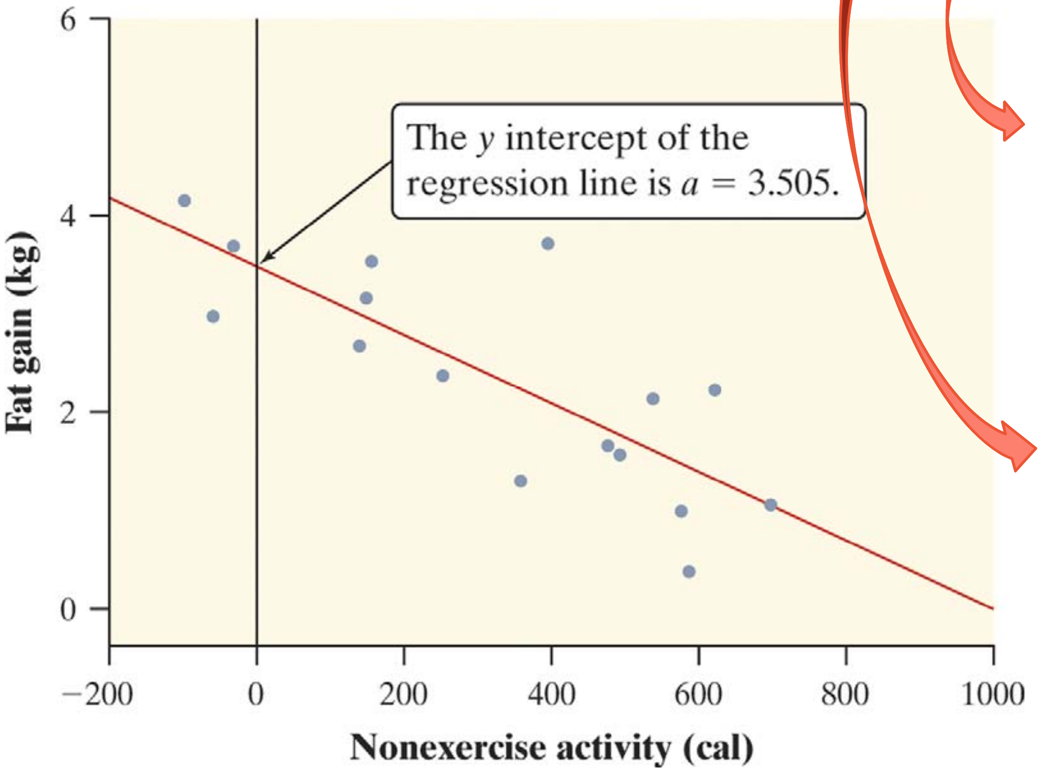
•*a* is the **y intercept**, the predicted value of *y* when $x = 0$.

Least-Squares Regression

Interpreting a Regression Line

Consider the regression line from the example “Does Fidgeting Keep You Slim?” Identify the slope and *y*-intercept and interpret each value in context.

$$\widehat{fatgain} = 3.505 - 0.00344(NEA\ change)$$



The *y* intercept of the regression line is *a* = 3.505.

The slope *b* = -0.00344 tells us that the amount of fat gained is predicted to go down by 0.00344 kg for each added calorie of NEA.

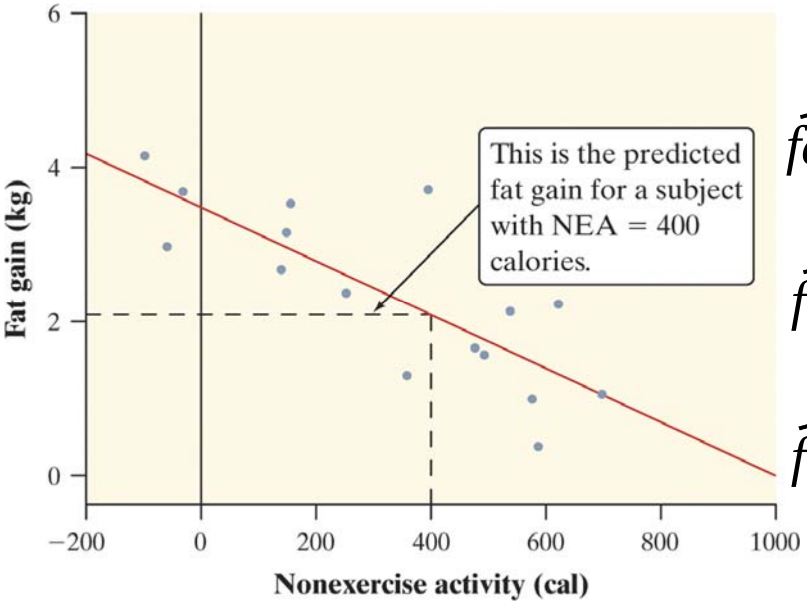
The *y*-intercept *a* = 3.505 kg is the fat gain estimated by this model if NEA does not change when a person overeats.

Least-Squares Regression

Prediction

We can use a regression line to predict the response $\hat{y}$ for a specific value of the explanatory variable *x*.

Use the NEA and fat gain regression line to predict the fat gain for a person whose NEA increases by 400 cal when she overeats.



$$\widehat{fatgain} = 3.505 - 0.00344(NEA\ change)$$

$$\widehat{fatgain} = 3.505 - 0.00344(400)$$

$$\widehat{fatgain} = 2.13$$

We predict a fat gain of 2.13 kg when a person with NEA = 400 calories.

- **Extrapolation**

We can use a regression line to predict the response *ŷ* for a specific value of the explanatory variable *x*. The accuracy of the prediction depends on how much the data scatter about the line.

While we can substitute any value of *x* into the equation of the regression line, we must exercise caution in making predictions outside the observed values of *x*.

Definition:

Extrapolation is the use of a regression line for prediction far outside the interval of values of the explanatory variable *x* used to obtain the line. Such predictions are often not accurate.

Don't make predictions using values of x that are much larger or much smaller than those that actually appear in your data.

Residuals

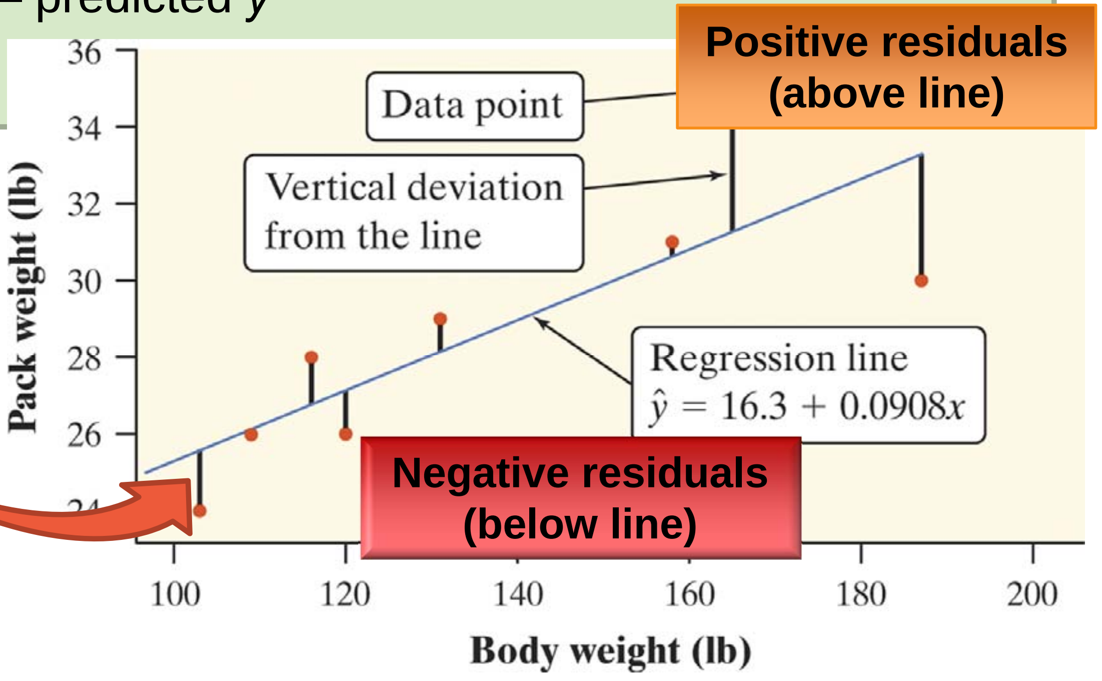
In most cases, no line will pass exactly through all the points in a scatterplot. A good regression line makes the vertical distances of the points from the line as small as possible.

Definition:

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

residual = observed *y* – predicted *y*

residual = y - $\hat{y}$



Positive residuals (above line)

Negative residuals (below line)

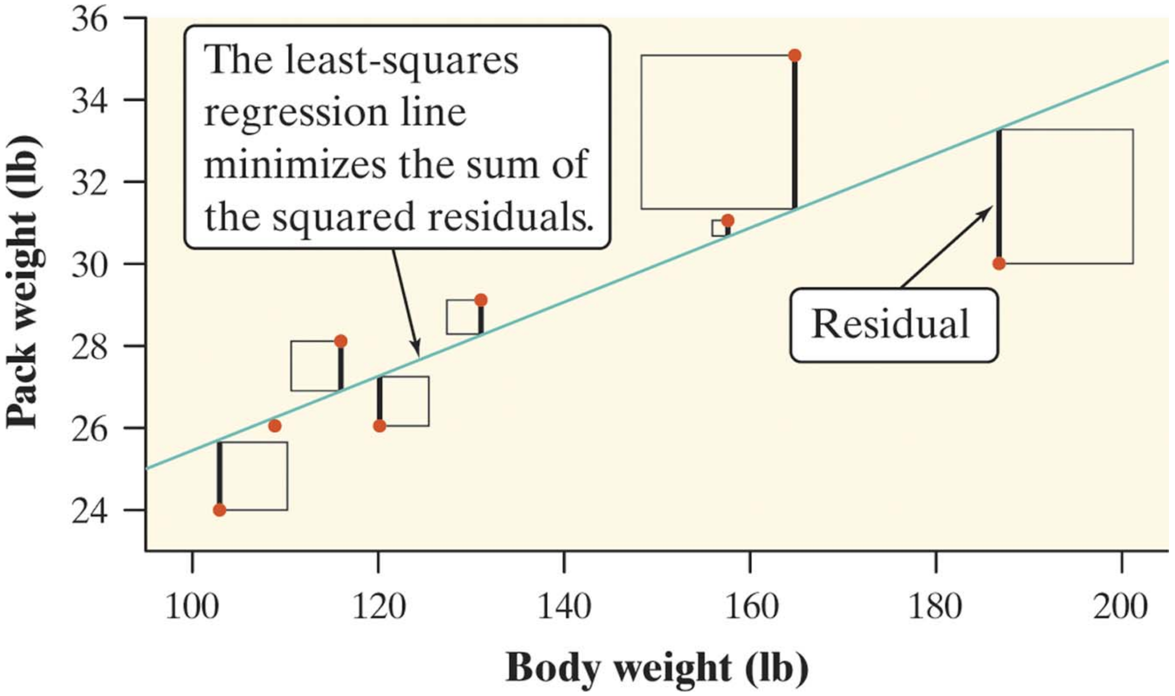
residual

Least-Squares Regression Line

Different regression lines produce different residuals. The regression line we want is the one that minimizes the sum of the squared residuals.

Definition:

The **least-squares regression line** of *y* on *x* is the line that makes the sum of the squared residuals as small as possible.



Least-Squares Regression Line

We can use technology to find the equation of the least-squares regression line. We can also write it in terms of the means and standard deviations of the two variables and their correlation.

Definition: **Equation of the least-squares regression line**

We have data on an explanatory variable *x* and a response variable *y* for *n* individuals. From the data, calculate the means and standard deviations of the two variables and their correlation. The least squares regression line is the line $\hat{y} = a + bx$ with

slope

$$b = r\frac{s_y}{s_x}$$

and ***y* intercept**

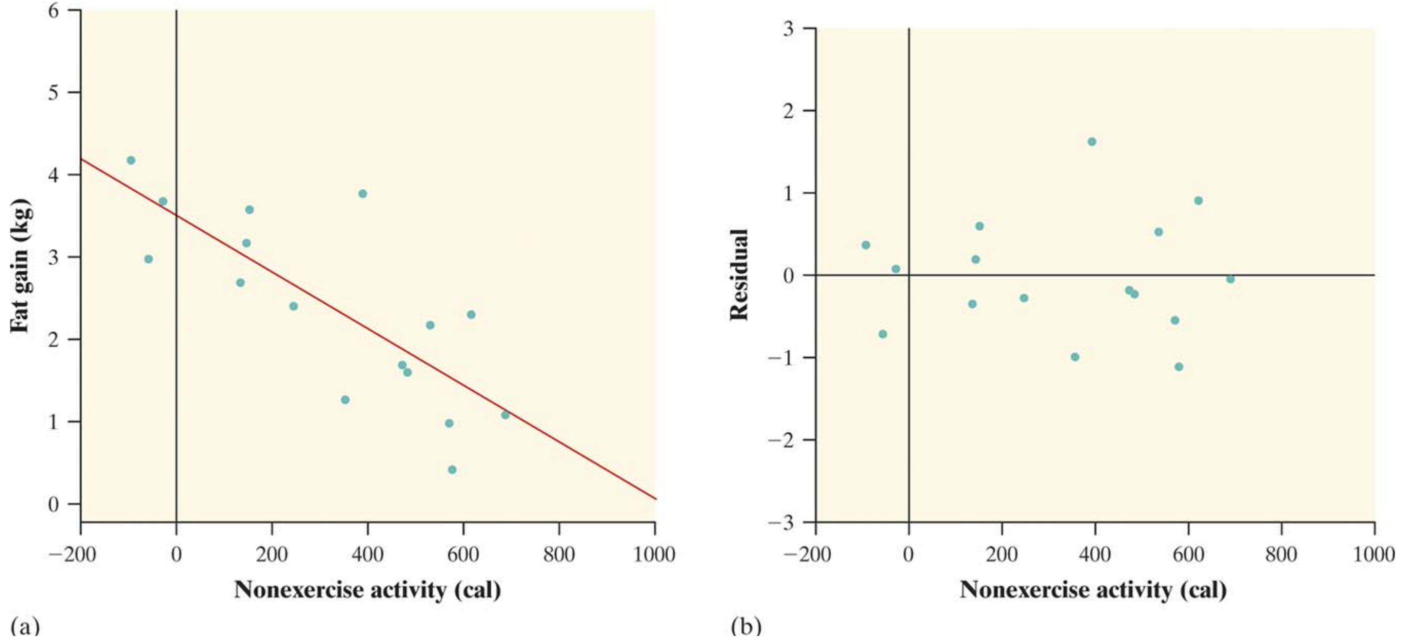
$$a = \bar{y} - b\bar{x}$$

Residual Plots

One of the first principles of data analysis is to look for an overall pattern and for striking departures from the pattern. A regression line describes the overall pattern of a linear relationship between two variables. We see departures from this pattern by looking at the residuals.

Definition:

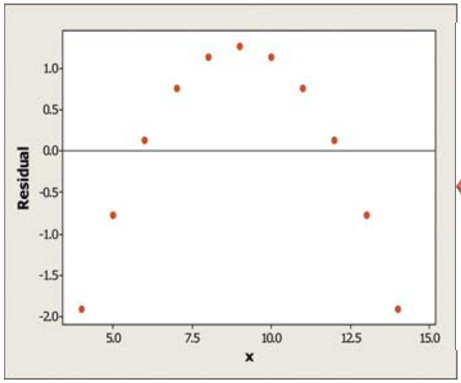
A **residual plot** is a scatterplot of the residuals against the explanatory variable. Residual plots help us assess how well a regression line fits the data.



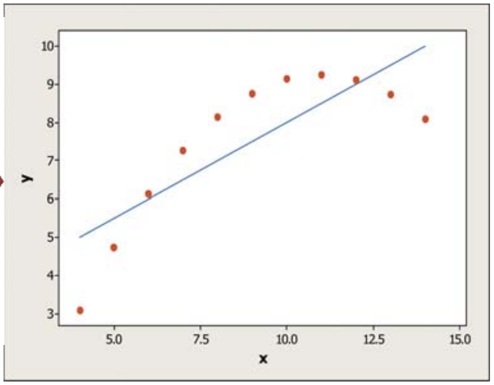
Interpreting Residual Plots

A residual plot magnifies the deviations of the points from the line, making it easier to see unusual observations and patterns.

1) The residual plot should show no obvious patterns
2) The residuals should be relatively small in size.



Pattern in residuals
Linear model not appropriate



Definition:

If we use a least-squares regression line to predict the values of a response variable *y* from an explanatory variable *x*, the **standard deviation of the residuals (s)** is given by

$$s = \sqrt{\frac{\sum residuals^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

The Role of r^2 in Regression

The standard deviation of the residuals gives us a numerical estimate of the average size of our prediction errors. There is another numerical quantity that tells us how well the least-squares regression line predicts values of the response *y*.

Definition:

The **coefficient of determination r^2** is the fraction of the variation in the values of *y* that is accounted for by the least-squares regression line of *y* on *x*. We can calculate r^2 using the following formula:

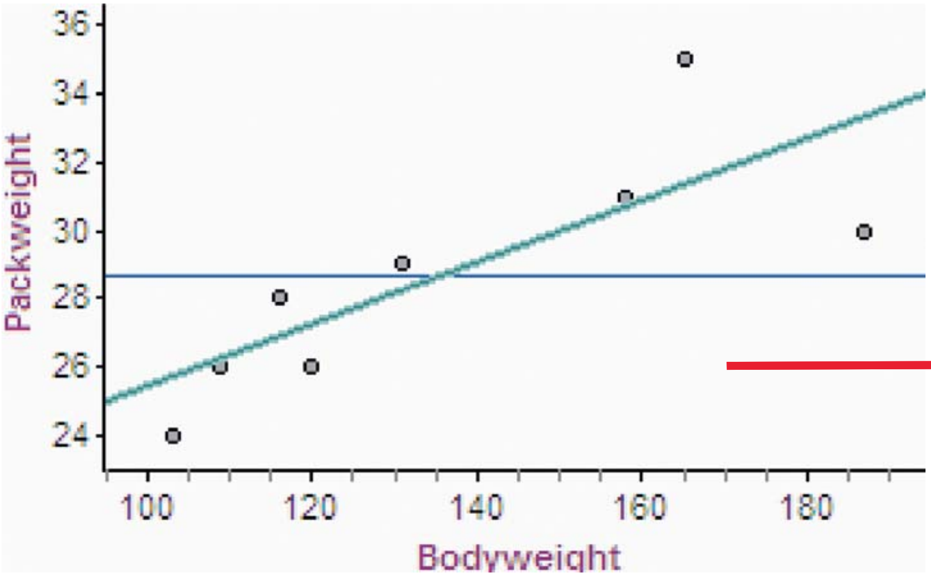
$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where $$SSE = \sum \text{residual}^2$$

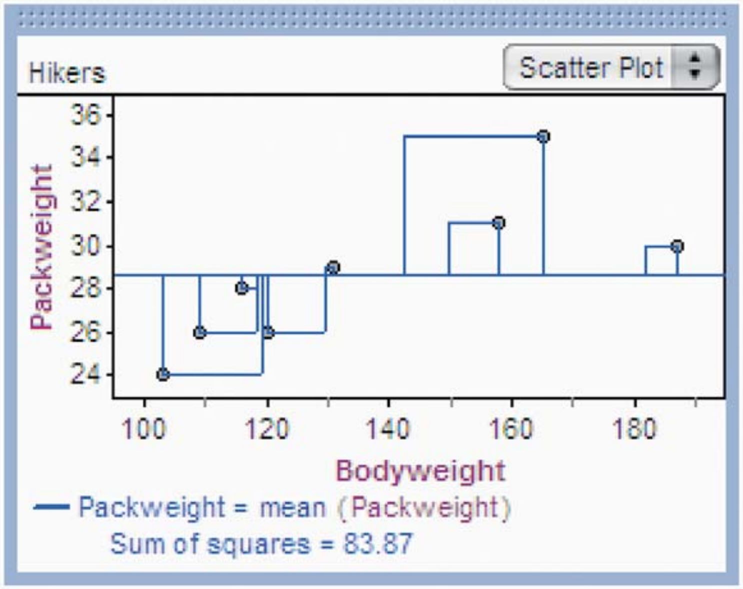
and $$SST = \sum (y_i - \bar{y})^2$$

The Role of r^2 in Regression

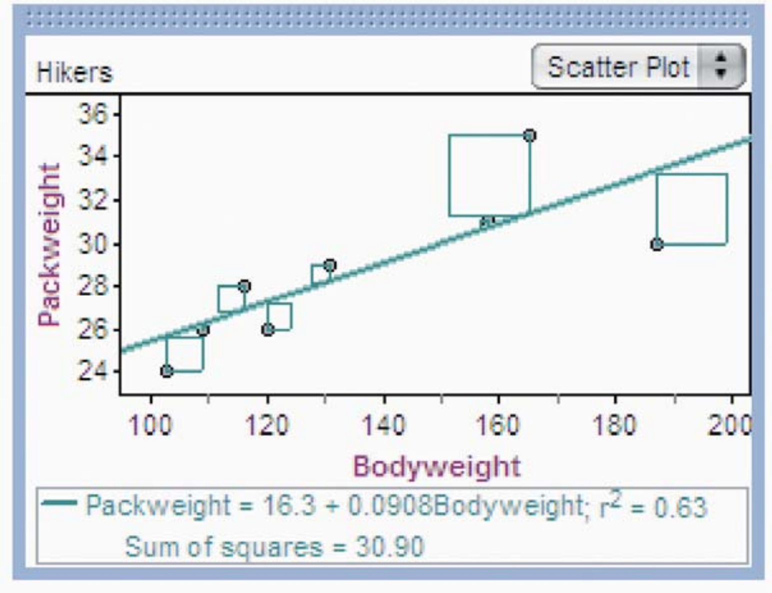
r^2 tells us how much better the LSRL does at predicting values of *y* than simply guessing the mean *y* for each value in the dataset. Consider the example on page 179. If we needed to predict a backpack weight for a new hiker, but didn't know each hikers weight, we could use the average backpack weight as our prediction.



The Role of r^2 in Regression (CONTINUED)



If we use the mean backpack weight as our prediction, the sum of the squared residuals is 83.87.
SST = 83.87



If we use the LSRL to make our predictions, the sum of the squared residuals is 30.90.
SSE = 30.90

SSE/SST = 30.97/83.87
SSE/SST = 0.368

Therefore, 36.8% of the variation in pack weight is *unaccounted for* by the least-squares regression line.

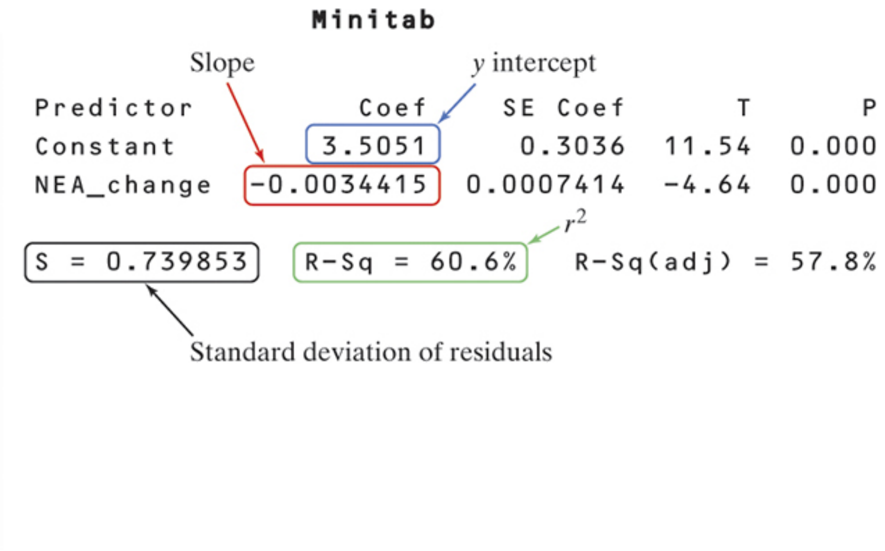
1 – SSE/SST = 1 – 30.97/83.87
r^2 = 0.632

63.2 % of the variation in backpack weight is accounted for by the linear model relating pack weight to body weight.

Interpreting Computer Regression Output

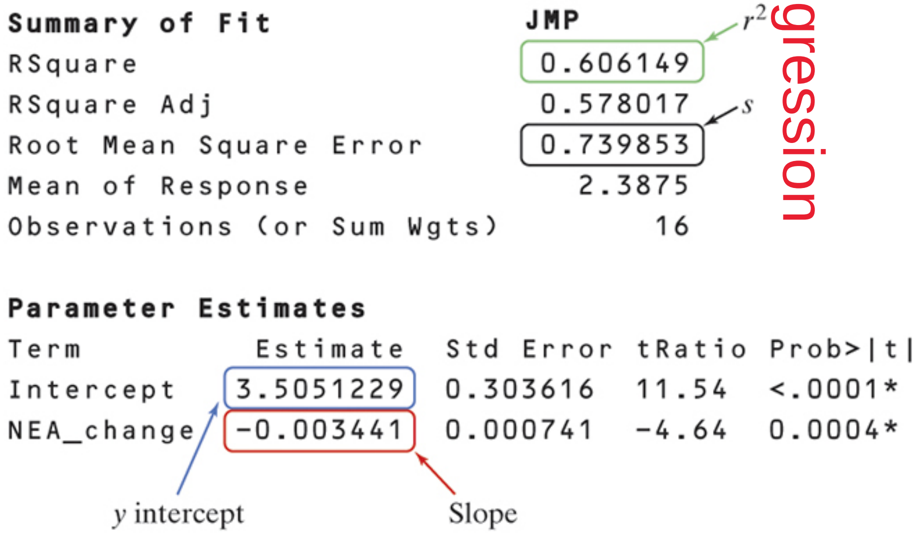
A number of statistical software packages produce similar regression output. Be sure you can locate

- the slope *b*,
- the *y* intercept *a*,
- and the values of *s* and r^2.



Minitab

```
Predictor        Coef      SE Coef       T      P
Constant       3.5051       0.3036   11.54  0.000
NEA_change  -0.0034415   0.0007414   -4.64  0.000

S = 0.739853   R-Sq = 60.6%    R-Sq(adj) = 57.8%
```

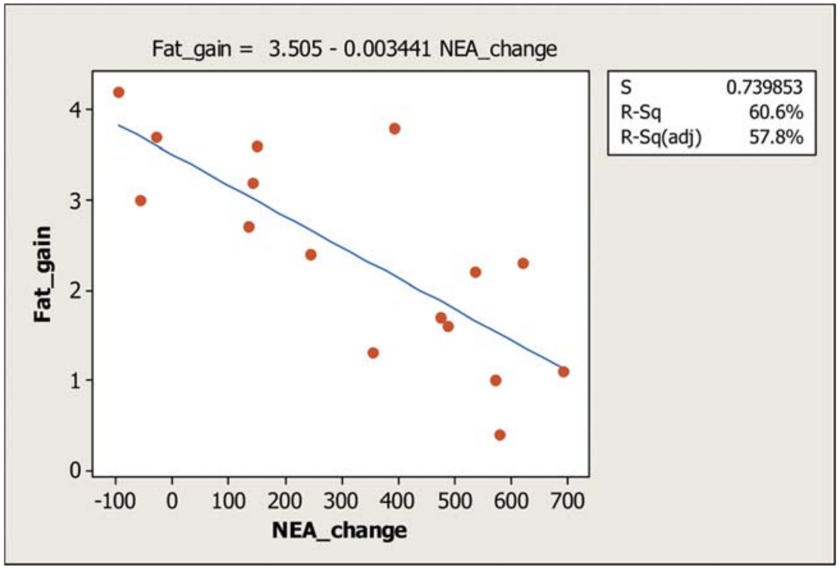


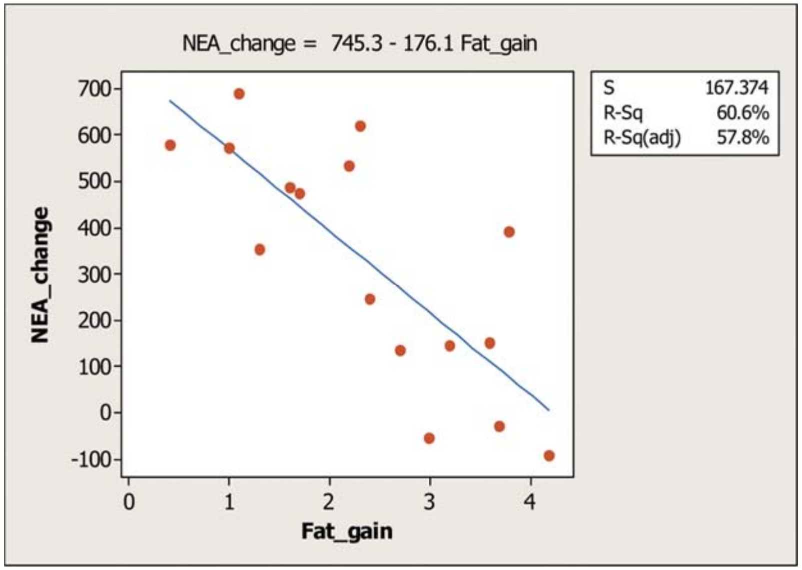
```
Summary of Fit                          JMP
RSquare                              0.606149
RSquare Adj                          0.578017
Root Mean Square Error               0.739853
Mean of Response                       2.3875
Observations (or Sum Wgts)                 16

Parameter Estimates
Term          Estimate   Std Error  tRatio  Prob>|t|
Intercept    3.5051229   0.303616   11.54   <.0001*
NEA_change   -0.003441   0.000741   -4.64   0.0004*
```

Correlation and Regression Wisdom

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, be aware of their limitations

1. The distinction between explanatory and response variables is important in regression.





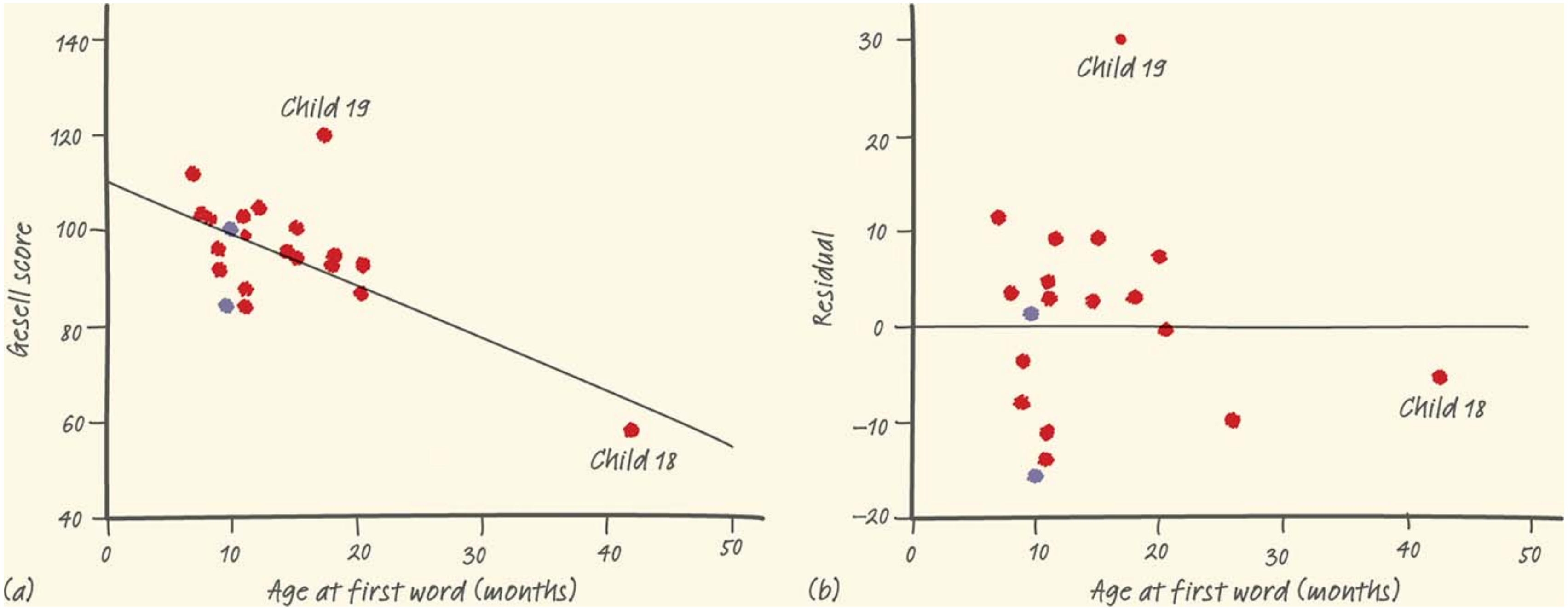
2. Correlation and regression lines describe only linear relationships.

Correlation and Regression Wisdom

3. Correlation and least-squares regression lines are not resistant.

Definition:

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the *y* direction but not the *x* direction of a scatterplot have large residuals. Other outliers may not have large residuals.

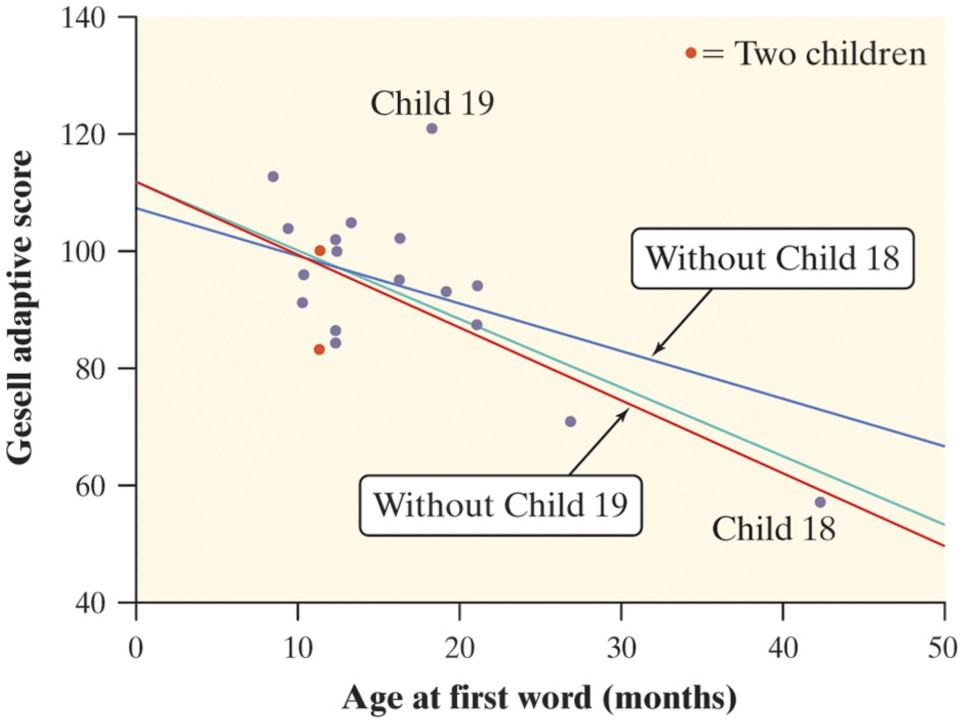


Correlation and Regression Wisdom

3. Correlation and least-squares regression lines are not resistant.

Definition:

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the *x* direction of a scatterplot are often influential for the least-squares regression line.



With all 19 children:
$r = -0.64$
$\hat{y} = 109.874 - 1.127x$

Without Child 19:
$r = -0.76$
$\hat{y} = 109.305 - 1.193x$

Without Child 18:
$r = -0.33$
$\hat{y} = 105.630 - 0.779x$

Correlation and Regression Wisdom

4. Association does not imply causation.

Association Does Not Imply Causation

An association between an explanatory variable *x* and a response variable *y*, even if it is very strong, is not by itself good evidence that changes in *x* actually cause changes in *y*.

A serious study once found that people with two cars live longer than people who only own one car. Owning three cars is even better, and so on. There is a substantial positive correlation between number of cars *x* and length of life *y*. Why?

Section 3.2
Least-Squares Regression

Summary

In this section, we learned that…

- ✓ 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 any value of *x*.
- ✓ The **slope *b*** of a regression line is the rate at which the predicted response *ŷ* changes along the line as the explanatory variable *x* changes. *b* is the *predicted* change in *y* when *x* increases by 1 unit.
- ✓ The ***y* intercept *a*** of a regression line is the predicted response for *ŷ* when the explanatory variable *x* = 0.
- ✓ Avoid **extrapolation**, predicting values outside the range of data from which the line was calculated.

Section 3.2
Least-Squares Regression

Summary

In this section, we learned that…

- ✓ The **least-squares regression line** is the straight line $\hat{y} = a + bx$ that minimizes the sum of the squares of the vertical distances of the observed points from the line.
- ✓ You can examine the fit of a regression line by studying the **residuals** (observed *y* – predicted *y*). Be on the lookout for points with unusually large residuals and also for nonlinear patterns and uneven variation in the **residual plot.**
- ✓ The **standard deviation of the residuals *s*** measures the average size of the prediction errors (residuals) when using the regression line.

Section 3.2
Least-Squares Regression

Summary

In this section, we learned that…

- ✓ The **coefficient of determination** r^2 is the fraction of the variation in one variable that is accounted for by least-squares regression on the other variable.
- ✓ Correlation and regression must be interpreted with caution. Plot the data to be sure the relationship is roughly linear and to detect **outliers** and **influential points**.
- ✓ Be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated.