# CHAPTER 3
# Describing Relationships

## 3.2
## Least-Squares Regression

The Practice of Statistics, 5th Edition
Starnes, Tabor, Yates, Moore

Bedford Freeman Worth Publishers

# Least-Squares Regression

## Learning Objectives

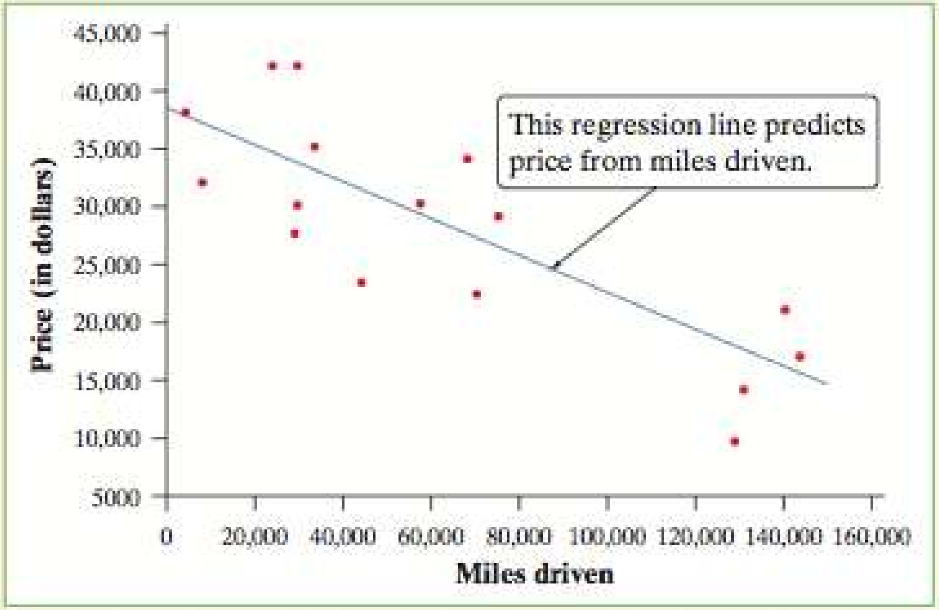
After this section, you should be able to:

- ✓ INTERPRET the slope and y intercept of a least-squares regression line.
- ✓ USE the least-squares regression line to predict *y* for a given *x*.
- ✓ CALCULATE and INTERPRET residuals and their standard deviation.
- ✓ EXPLAIN the concept of least squares.
- ✓ DETERMINE the equation of a least-squares regression line using a variety of methods.
- ✓ CONSTRUCT and INTERPRET residual plots to assess whether a linear model is appropriate.
- ✓ ASSESS how well the least-squares regression line models the relationship between two variables.
- ✓ DESCRIBE how the slope, *y* intercept, standard deviation of the residuals, and $r^2$ are influenced by outliers.

# 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.

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*.



# 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*.

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,

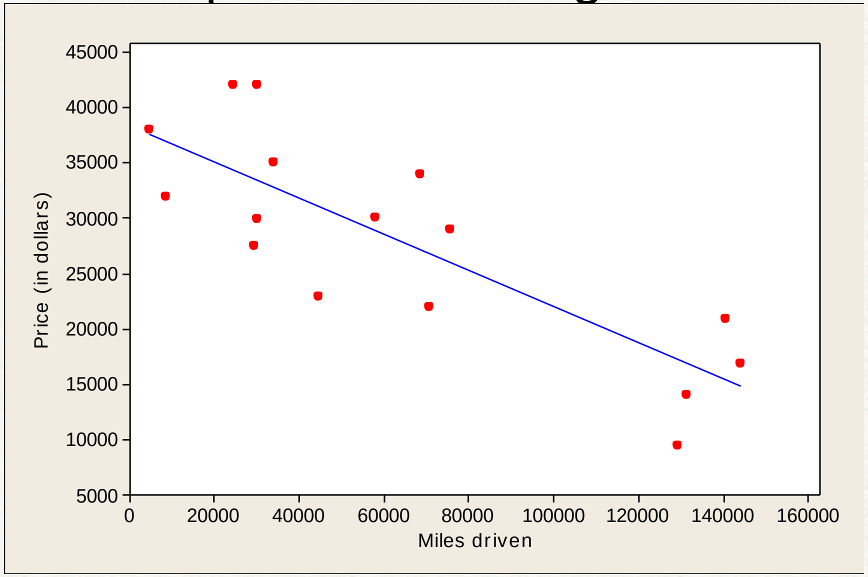
•$\hat{y}$ (read "y hat") is the **predicted value** of the response variable *y* for a given value of the explanatory variable *x*.

•*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 $x = 0$.

# Example: Interpreting slope and *y* intercept

The equation of the regression line shown is



$$\widehat{\text{price}} = 38257 - 0.1629(\text{miles driven})$$

PROBLEM: Identify the slope and y intercept of the regression line.

Interpret each value in context.

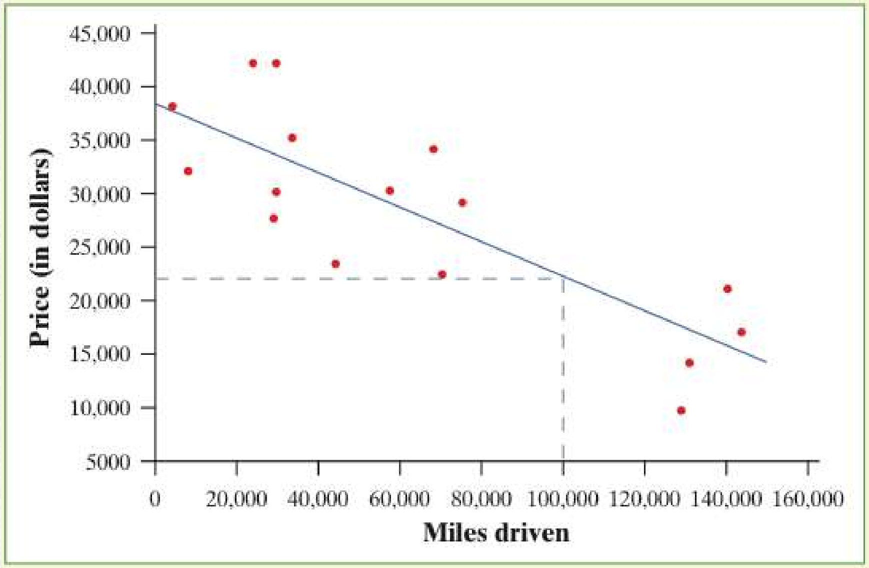
SOLUTION: The slope *b* = -0.1629 tells us that the price of a used Ford F-150 is predicted to go down by 0.1629 dollars (16.29 cents) for each additional mile that the truck has been driven.

The *y* intercept *a* = 38,257 is the predicted price of a Ford F-150 that has been driven 0 miles.

# Prediction

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

Use the regression line to predict price for a Ford F-150 with 100,000 miles driven.



$$\widehat{\text{price}} = 38257 - 0.1629(\text{miles driven})$$

$$\widehat{\text{price}} = 38,257 - 0.1629(100,000)$$

$$\widehat{\text{price}} = 21,967 \text{ dollars}$$

# Extrapolation

We can use a regression line to predict the response $\hat{y}$ 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*.

**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.*

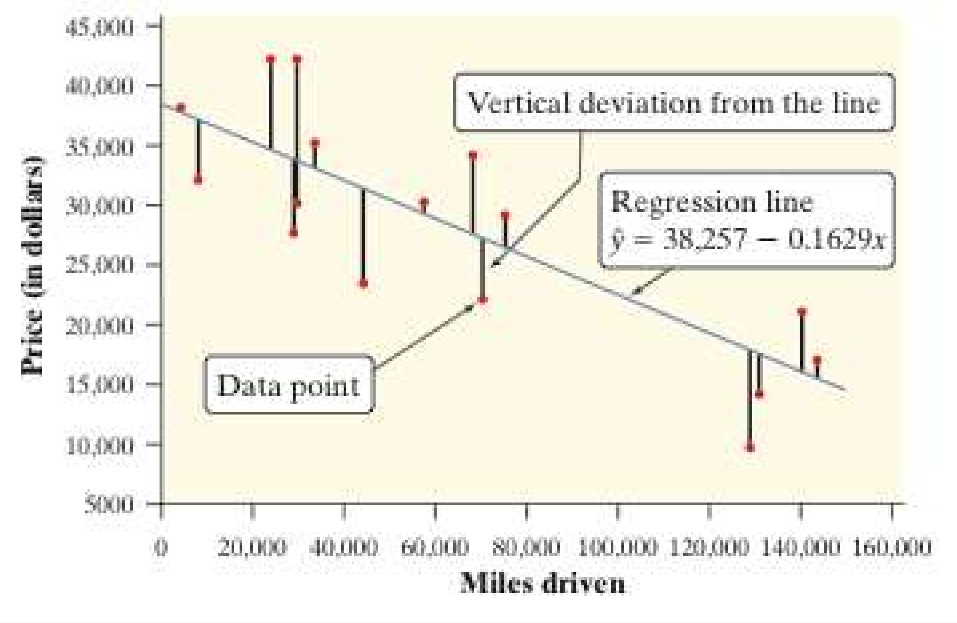
# Residuals

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.

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*

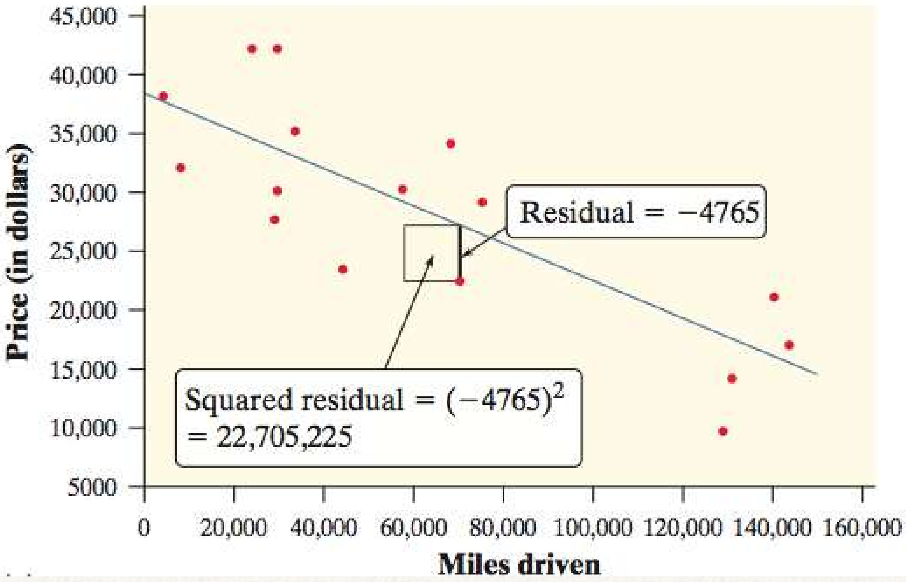
residual = $y - \hat{y}$



# Least Squares Regression Line

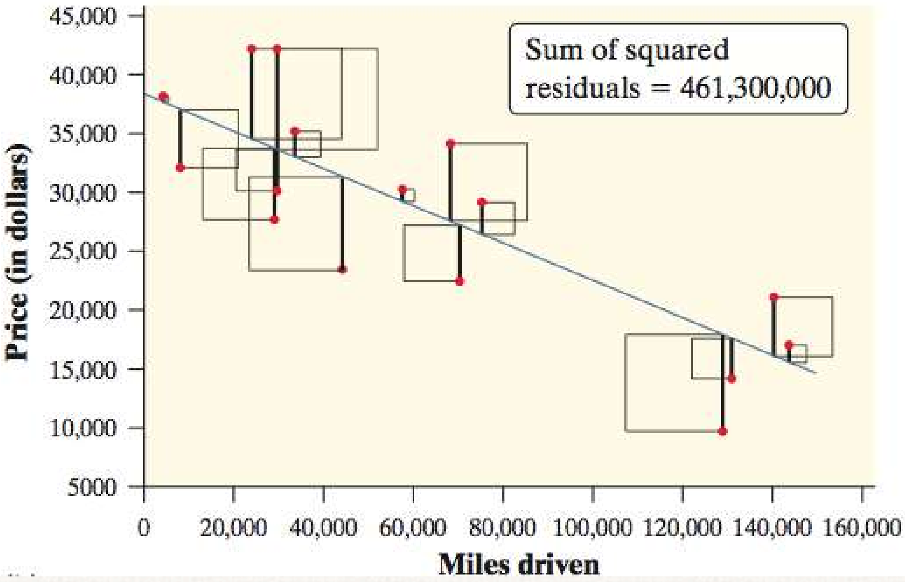
Different regression lines produce different residuals.  The regression line we want is the one that minimizes the sum of the squared residuals.

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



Residual = −4765

Squared residual = $(-4765)^2$
= 22,705,225

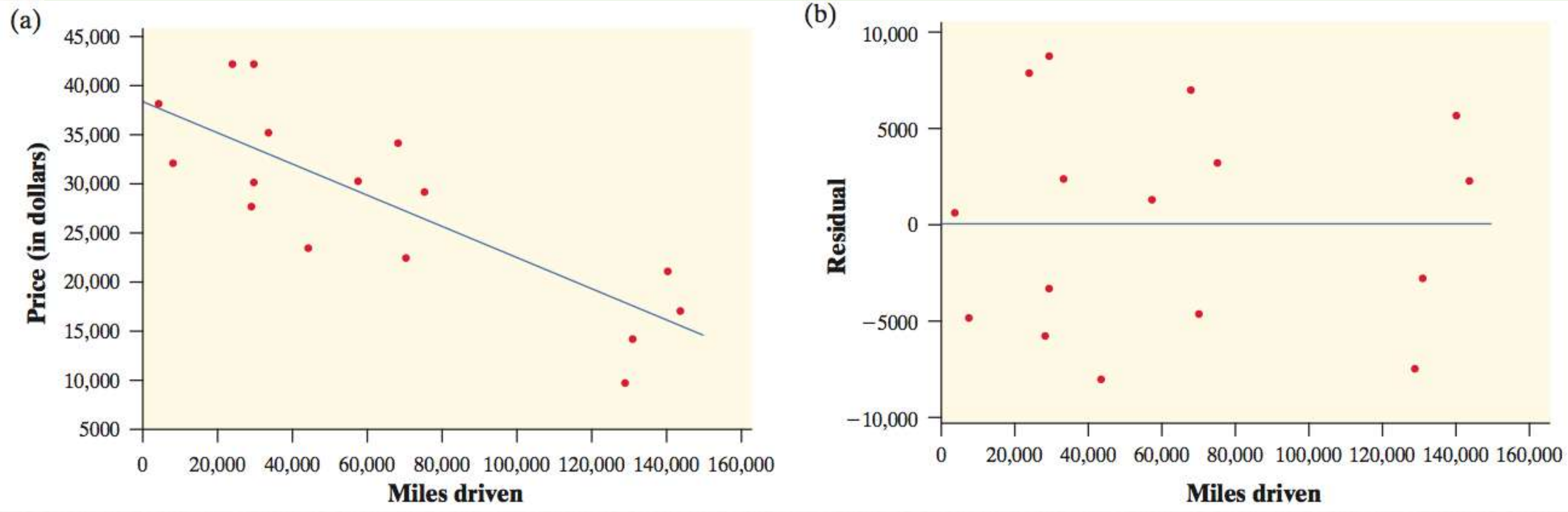


Sum of squared residuals = 461,300,000

# 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.

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.

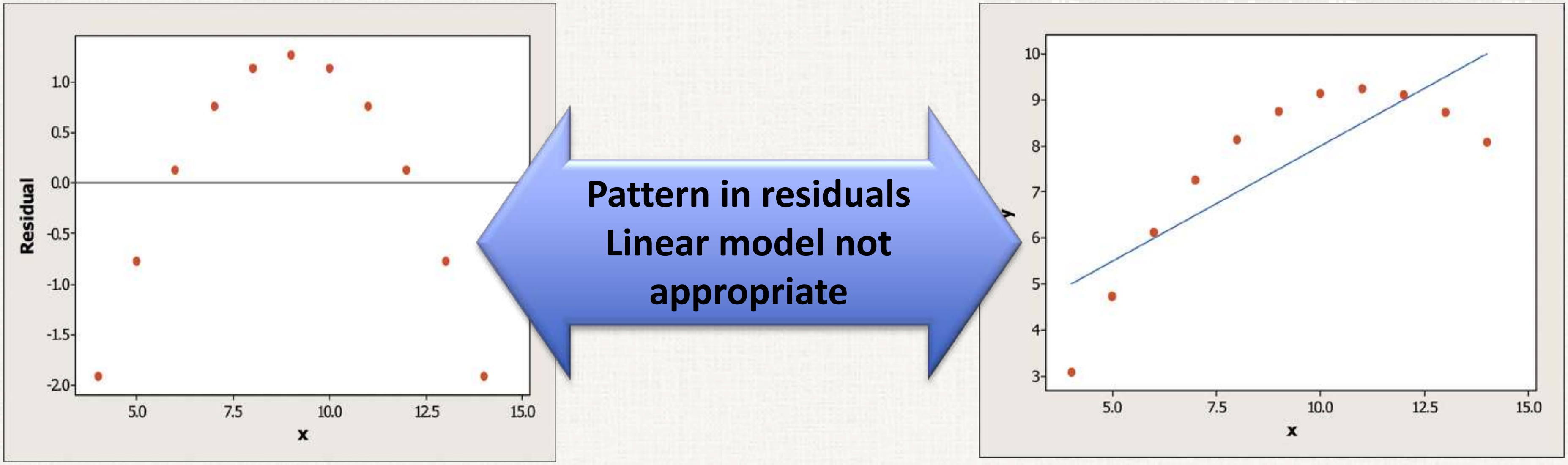


# Examining Residual Plots

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

The residual plot should show no obvious patterns

The residuals should be relatively small in size.



Pattern in residuals
Linear model not appropriate

# Standard Deviation of the Residuals

To assess how well the line fits all the data, we need to consider the residuals for each observation, not just one. Using these residuals, we can estimate the "typical" prediction error when using the least-squares regression line.

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

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

This value gives the approximate size of a "typical" prediction error (residual).

# The Coefficient of Determination

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*.

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:

$$r^2 = 1 - \frac{\sum \text{residuals}^2}{\sum (y_i - \bar{y})^2}$$

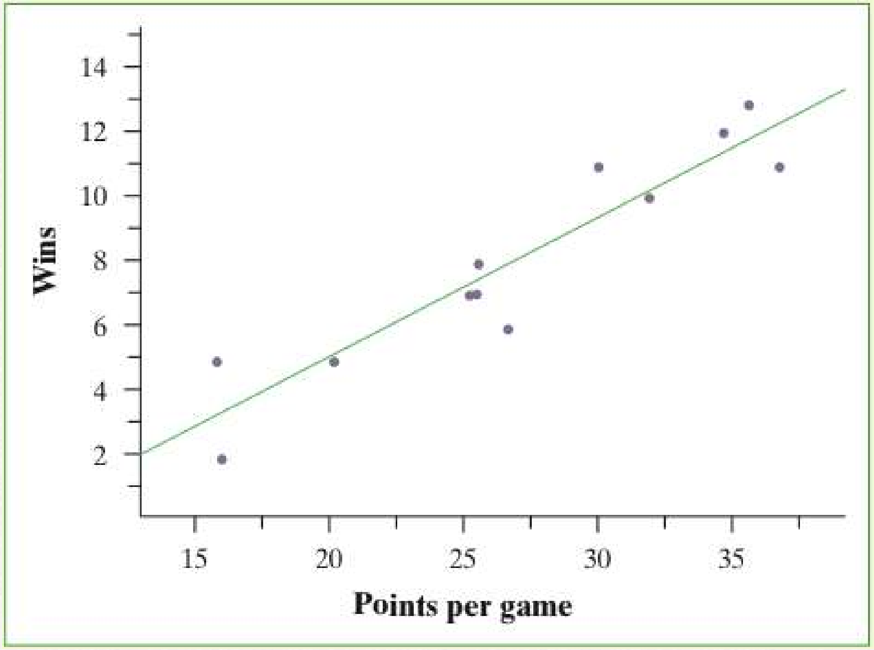
*r*² tells us how much better the LSRL does at predicting values of *y* than simply guessing the mean *y* for each value in the dataset.

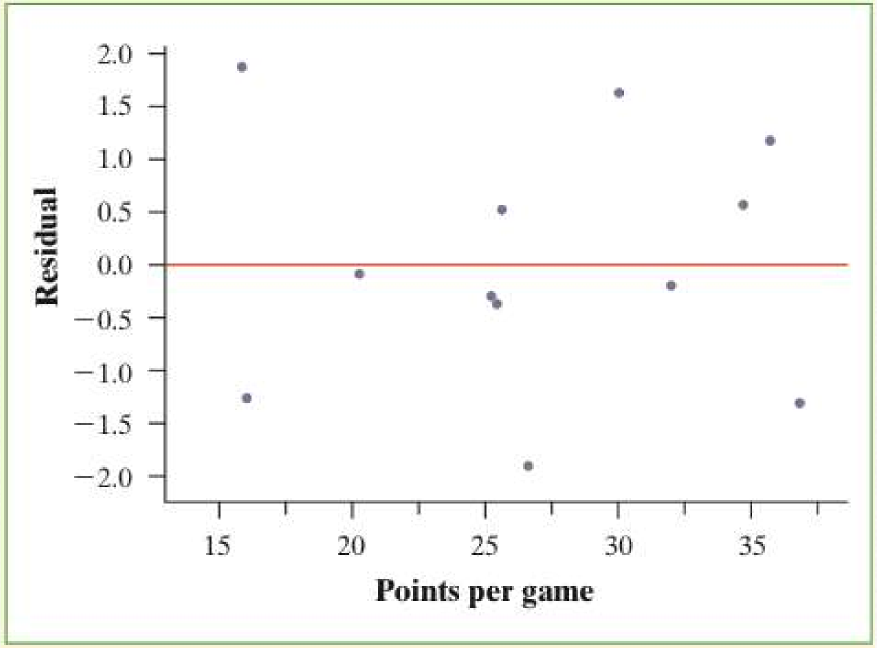
# Example: Residual plots, *s*, and $r^2$

In Section 3.1, we looked at the relationship between the average number of points scored per game *x* and the number of wins *y* for the 12 college football teams in the Southeastern Conference. A scatterplot with the least-squares regression line and a residual plot are shown.

The equation of the least-squares regression line is

$y$-hat = −3.75 + 0.437$x$.  Also, $s = 1.24$ and $r^2 = 0.88$.





# Example: Residual plots, *s*, and $r^2$

**(a) Calculate and interpret the residual for South Carolina, which scored 30.1 points per game and had 11 wins.**

The predicted amount of wins for South Carolina is

$$\hat{y} = -3.75 + 0.437(30.1) = 9.40 \text{ wins}$$

The residual for South Carolina is

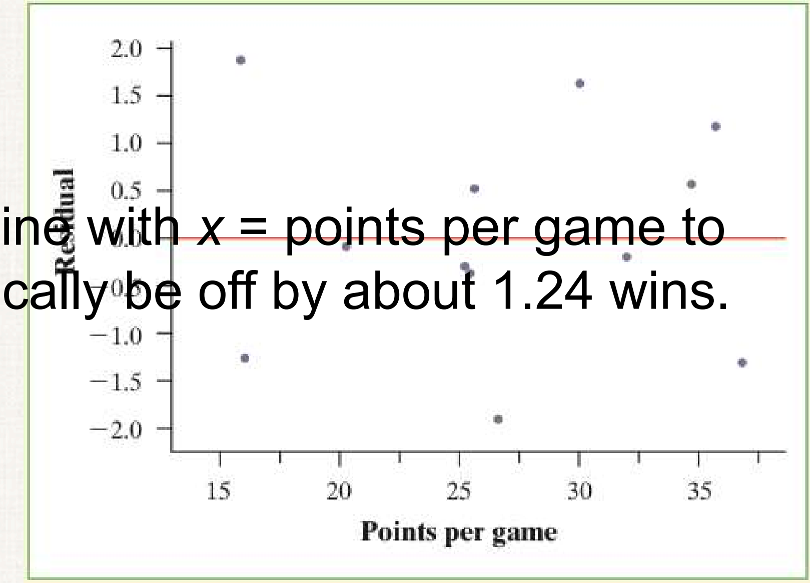
$$\text{residual} = y - \hat{y} = 11 - 9.40 = 1.60 \text{ wins}$$

South Carolina won 1.60 more games than expected, based on the number of points they scored per game.

# Example: Residual plots, *s*, and $r^2$

**(b) Is a linear model appropriate for these data? Explain.**

Because there is no obvious pattern left over in the residual plot, the linear model is appropriate.



**(c) Interpret the value *s* = 1.24.**

When using the least-squares regression line with *x* = points per game to predict *y* = the number of wins, we will typically be off by about 1.24 wins.

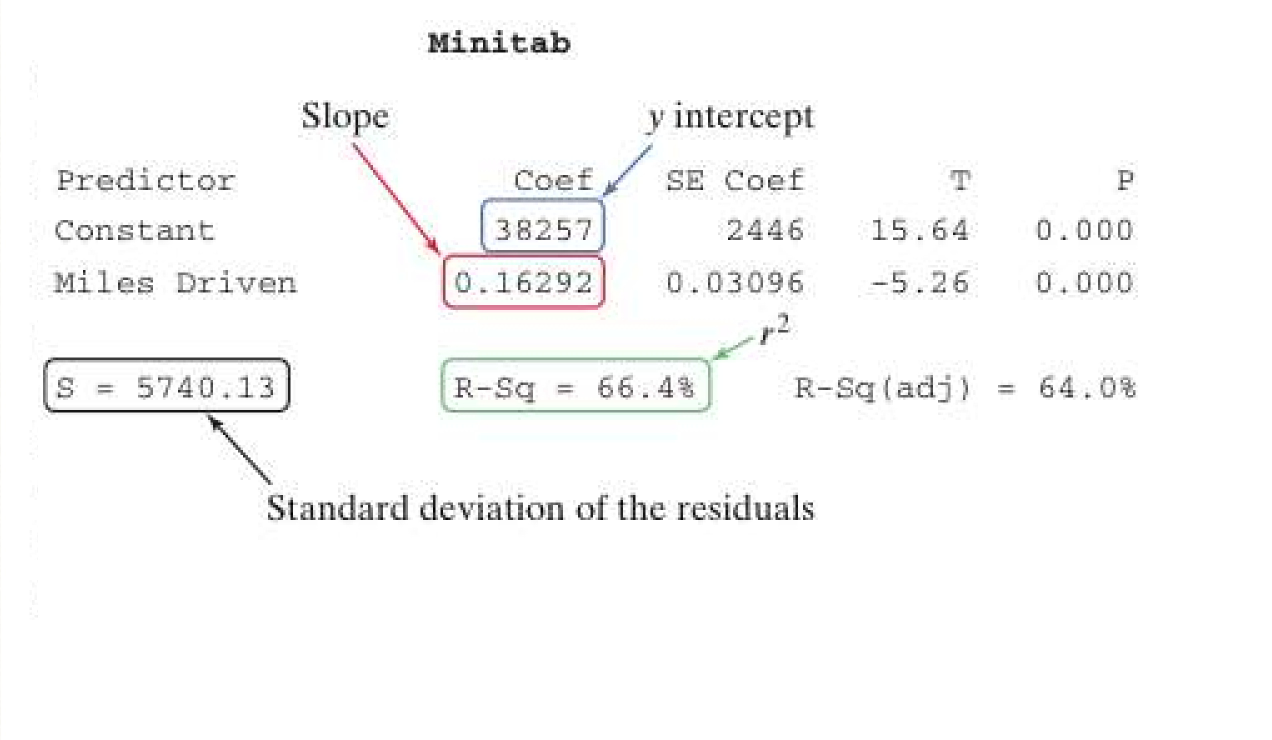
**(d) Interpret the value $r^2$ = 0.88.**

About 88% of the variation in wins is accounted for by the linear model relating wins to points per game.

# 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*
- the values of *s* and $r^2$



```
                 Minitab

Predictor            Coef    SE Coef       T       P
Constant            38257       2446   15.64   0.000
Miles Driven      0.16292    0.03096   -5.26   0.000

S = 5740.13          R-Sq = 66.4%     R-Sq(adj) = 64.0%
```

# Regression to the Mean

Using technology is often the most convenient way to find the equation of a least-squares regression line. It is also possible to calculate the equation of the least- squares regression line using only the means and standard deviations of the two variables and their correlation.

**How to Calculate 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 the standard deviations of the two variables and their correlation *r*.

The least-squares regression line is the line $\hat{y} = a + bx$ with **slope**

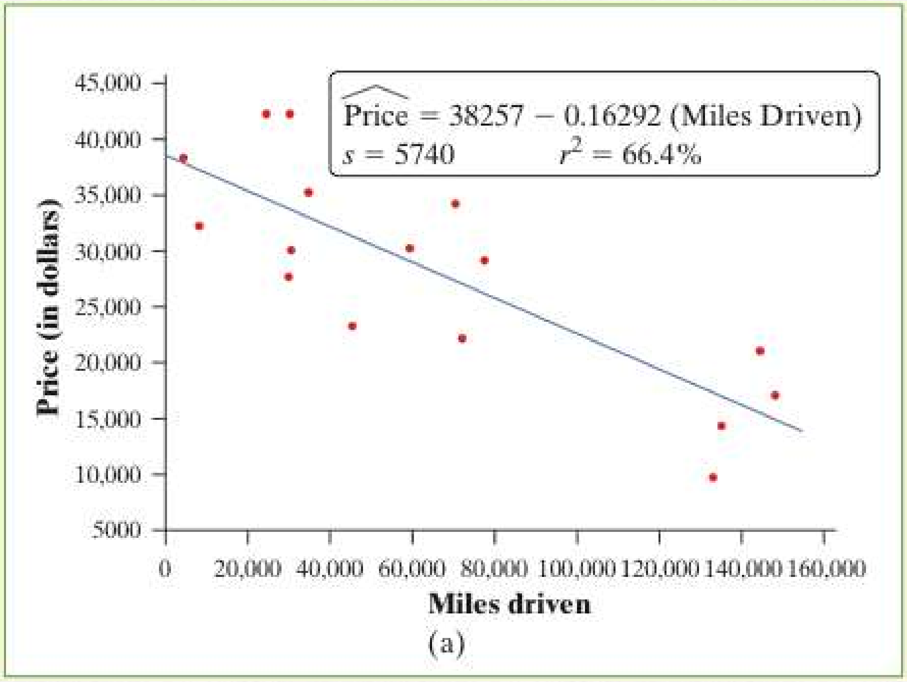
$$b = r\frac{s_y}{s_x}$$

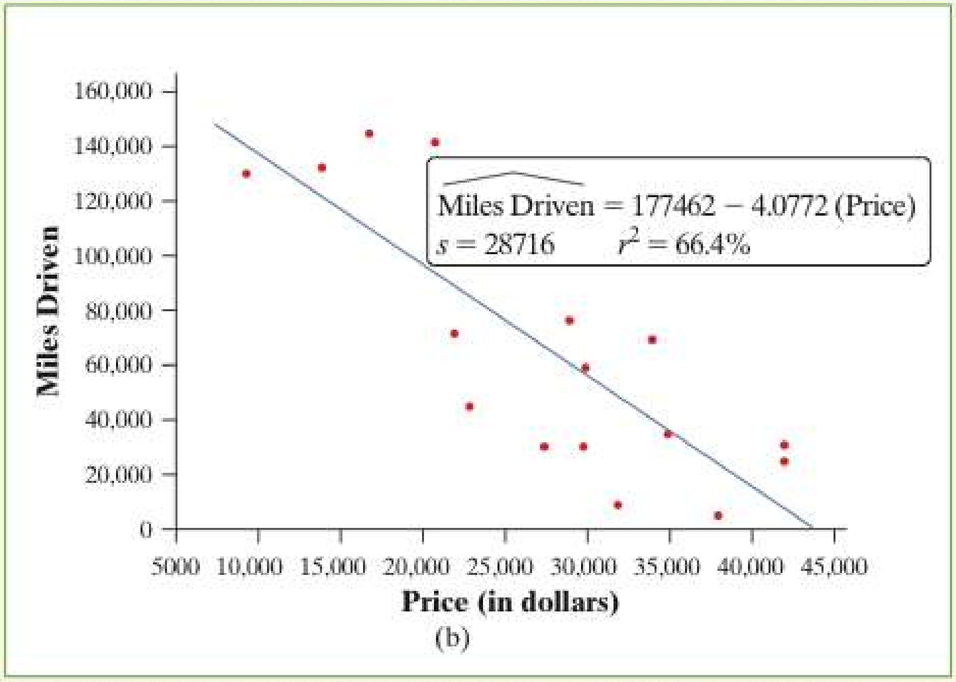
And ***y* intercept** $a = \bar{y} - b\bar{x}$

# 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.

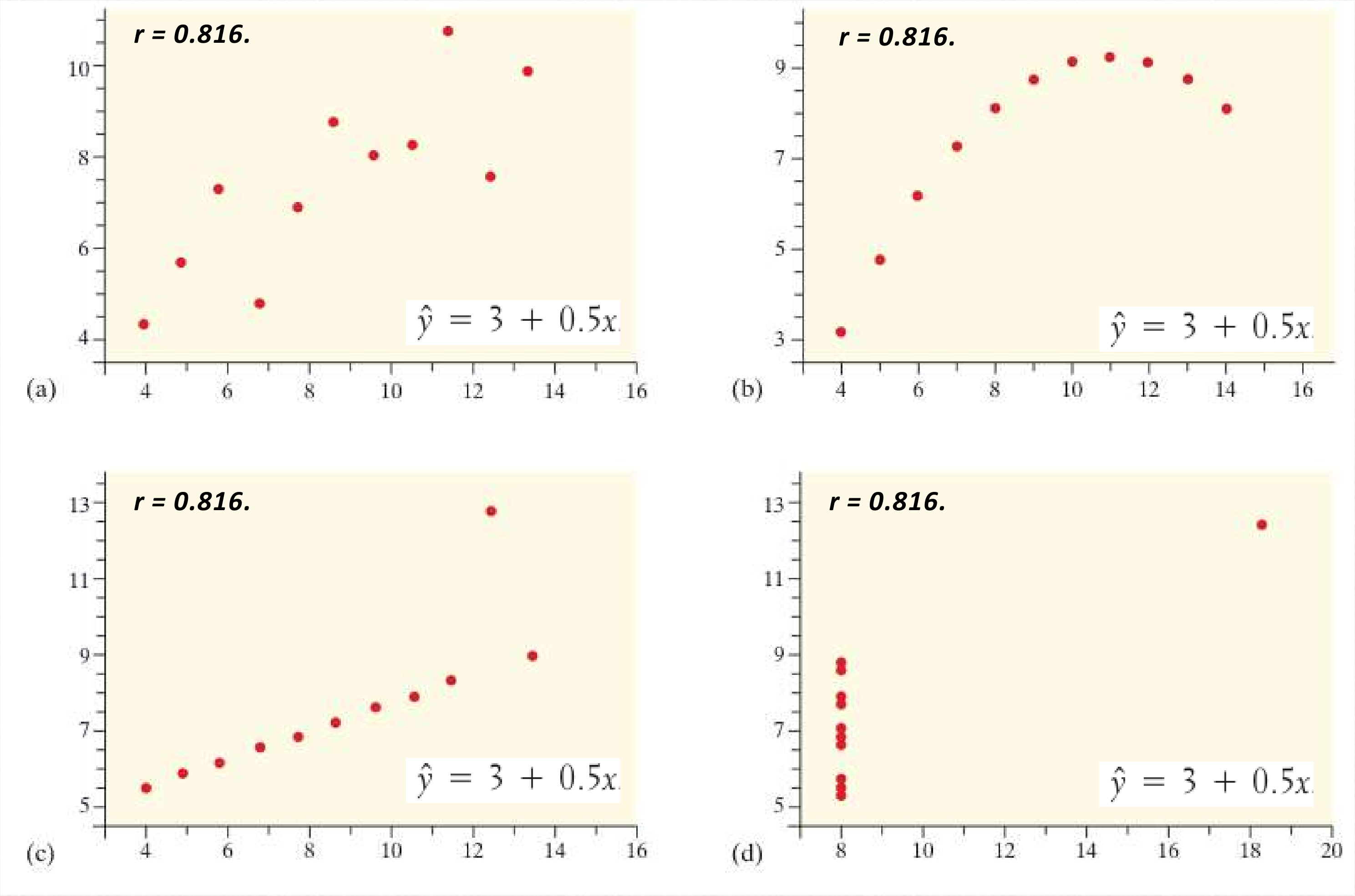
1. The distinction between explanatory and response variables is important in regression.





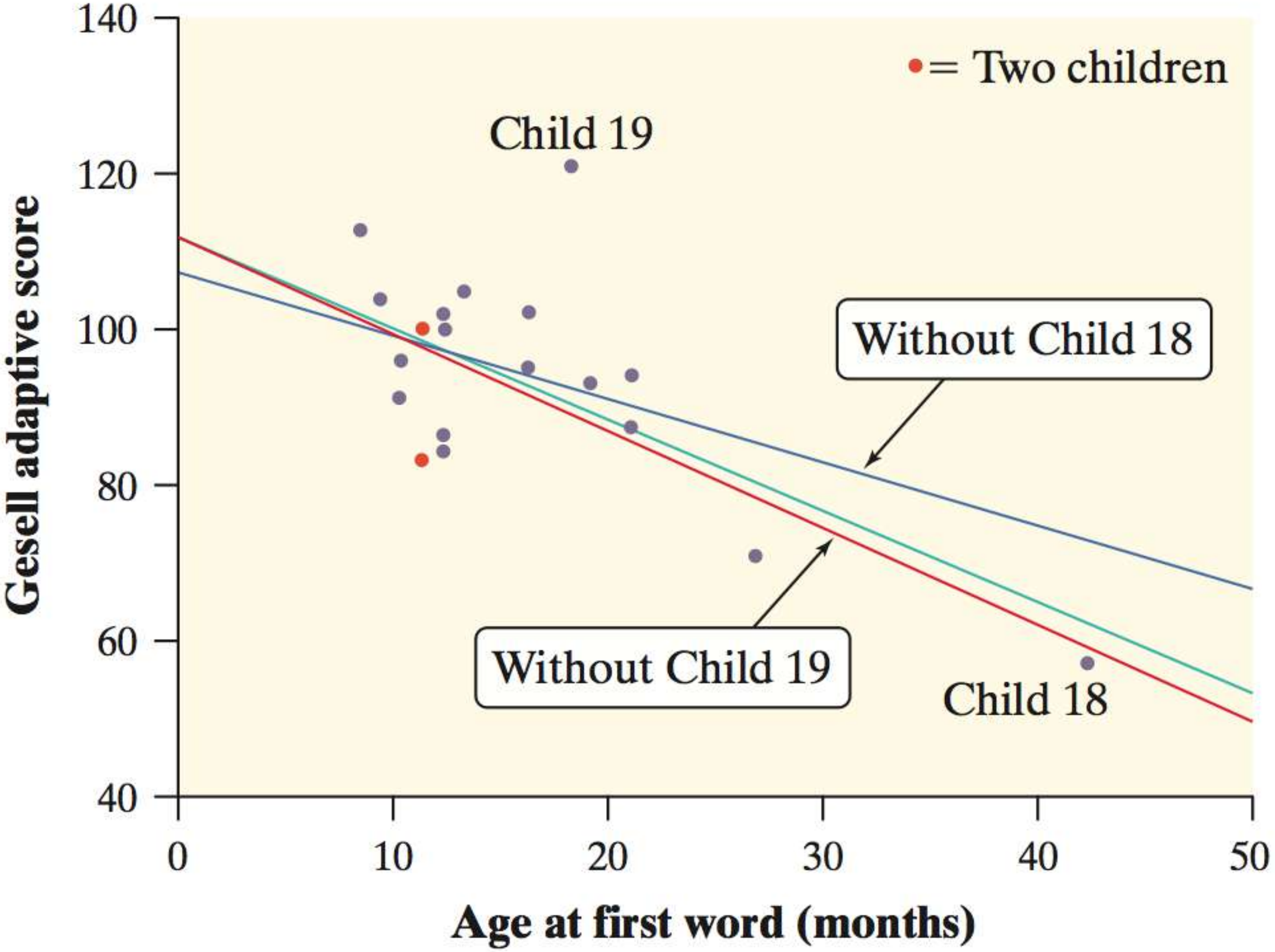
# Correlation and Regression Wisdom

**2**. Correlation and regression lines describe only linear relationships.



# Correlation and Regression Wisdom

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



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:

$r = -0.33$

$\hat{y} = 105.630 - 0.779x$

# Outliers and Influential Observations in Regression

Least-squares lines make the sum of the squares of the vertical distances to the points as small as possible. A point that is extreme in the *x* direction with no other points near it pulls the line toward itself. We call such points **influential**.

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.

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.

# Least-Squares Regression

## Section Summary

In this section, we learned how to…

- ✓ INTERPRET the slope and y intercept of a least-squares regression line.
- ✓ USE the least-squares regression line to predict y for a given x.
- ✓ CALCULATE and INTERPRET residuals and their standard deviation.
- ✓ EXPLAIN the concept of least squares.
- ✓ DETERMINE the equation of a least-squares regression line using a variety of methods.
- ✓ CONSTRUCT and INTERPRET residual plots to assess whether a linear model is appropriate.
- ✓ ASSESS how well the least-squares regression line models the relationship between two variables.
- ✓ DESCRIBE how the slope, y intercept, standard deviation of the residuals, and $r^2$ are influenced by outliers.