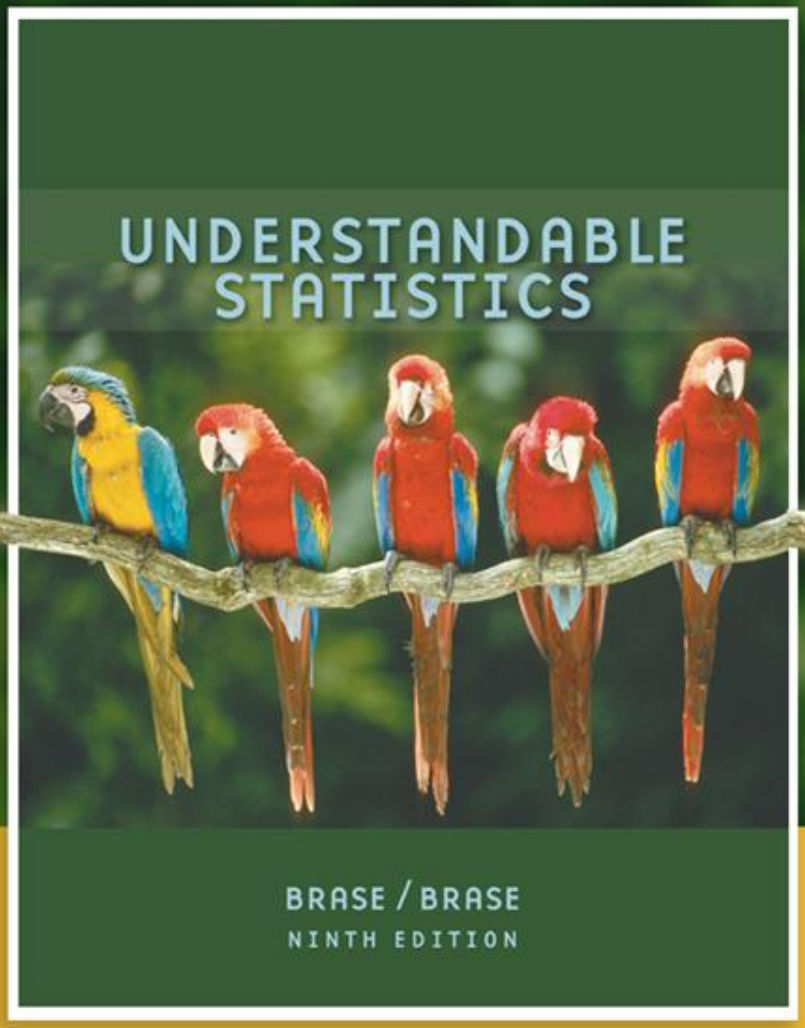


Chapter 10

Regression

Understandable Statistics
Ninth Edition

By Brase and Brase
Prepared by Yixun Shi
Bloomsburg University of Pennsylvania

Scatter Diagrams

- A graph in which pairs of points, *(x, y),* are plotted with *x* on the horizontal axis and *y* on the vertical axis.
- The explanatory variable is *x*.
- The response variable is *y*.
- One goal of plotting paired data is to determine if there is a linear relationship between *x* and *y*.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Finding the Best Fitting Line

- One goal of data analysis is to find a mathematical equation that “best” represents the data.

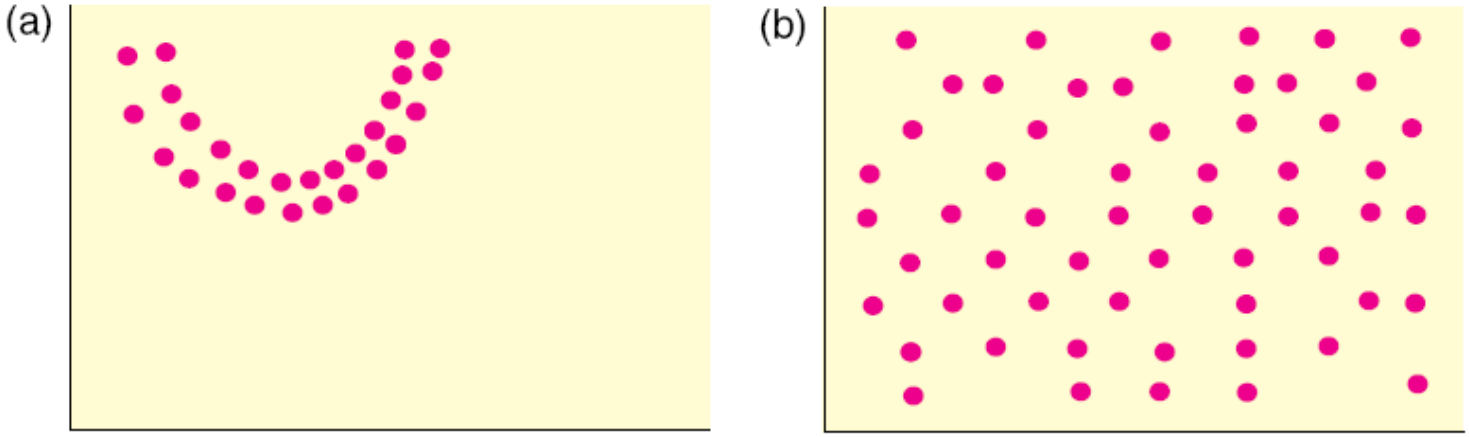
- For our purposes, “best” means the line that comes closest to each point on the scatter diagram.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Not All Relationships are Linear

FIGURE 10-2

Scatter Diagrams with No Linear Correlation



Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Correlation Strength

- Correlation is a measure of the *linear* relationship between two quantitative variables.
- If the points lie exactly on a straight line, we say there is *perfect correlation*.
- As the strength of the relationship decreases, the correlation moves from *perfect* to *strong* to *moderate* to *weak*.

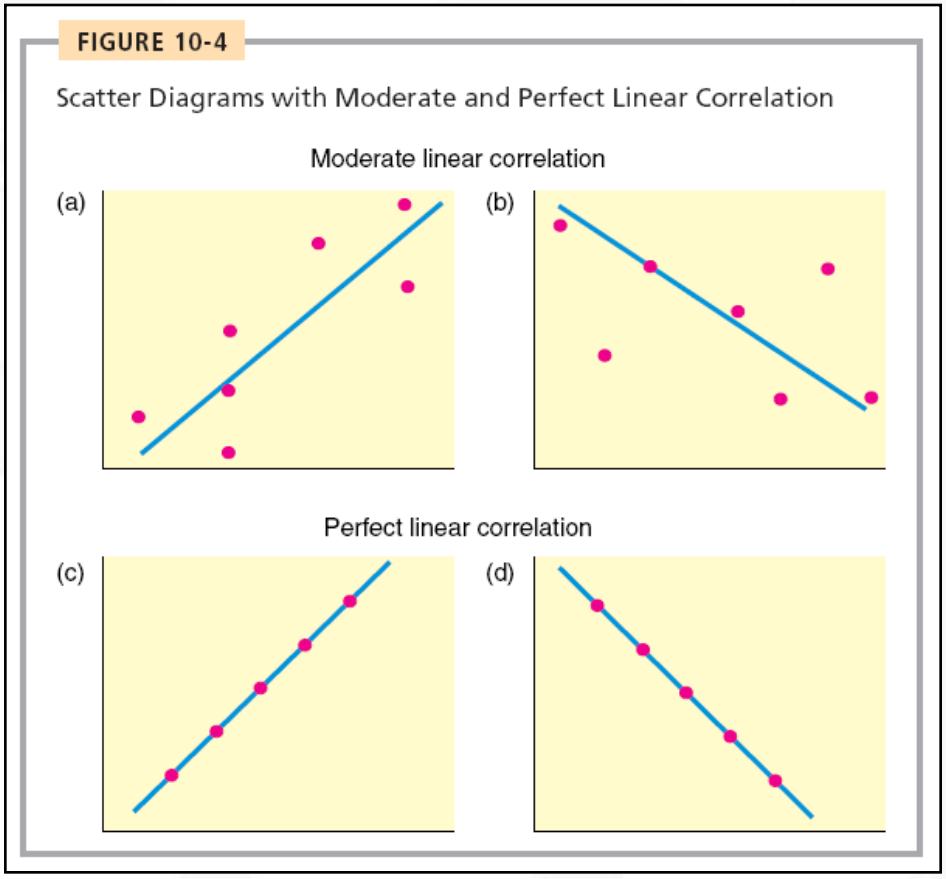
Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Correlation

- If the response variable, *y*, tends to increase as the explanatory variable, *x*, increases, the variables have *positive correlation*.
- If the response variable, *y*, tends to decrease as the explanatory variable, *x*, increases, the variables have *negative correlation*.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Correlation Examples



Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Correlation Coefficient

- The mathematical measurement that describes the correlation is called the *sample correlation coefficient.*

- Denoted by *r*.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

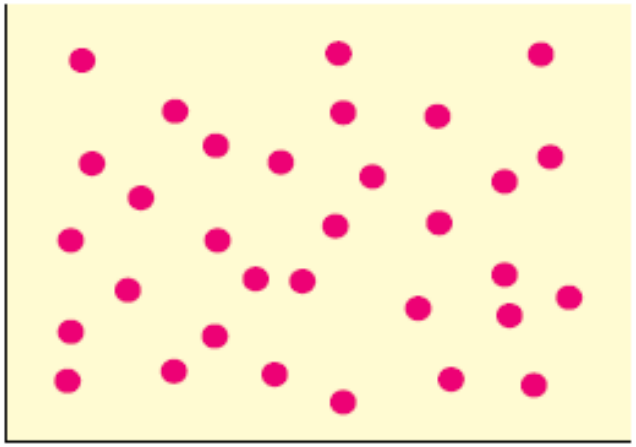
Correlation Coefficient Features

1) *r* has no units.
2) $-1 \leq r \leq 1$
3) Positive values of r indicate a positive relationship between *x* and *y* (likewise for negative values).
4) $r = 0$ indicates no *linear* relationship.
5) Switching the explanatory variable and response variable does not change *r.*
6) Changing the units of the variables does not change *r.*

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

$r = 0$

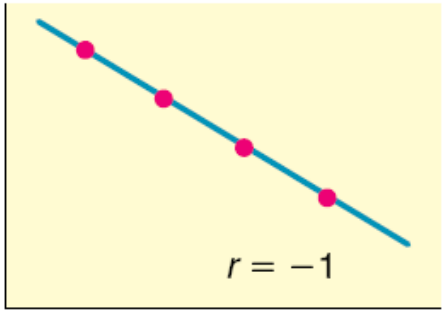
There is no linear relation for the points of the scatter diagram.

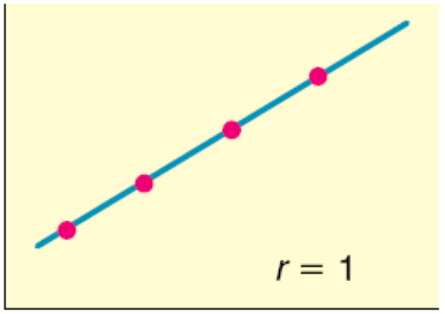


Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

$r = 1$ or $r = -1$

There is a perfect linear relation between x and y values; all points lie on the least-squares line.

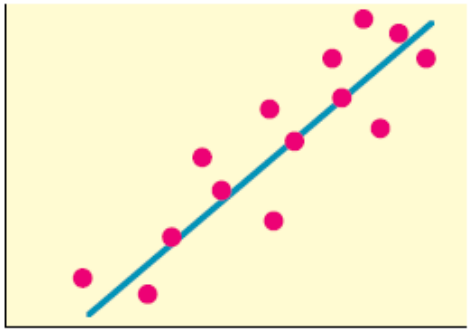




Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

$0 < r < 1$

The x and y values have a *positive correlation.* By this, we mean that *large* x values are associated with *large* y values, and *small* x values are associated with *small* y values.

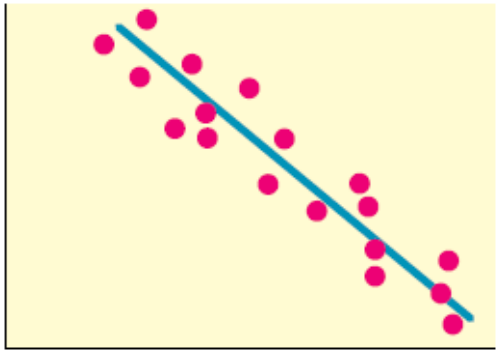


As we go from left to right, the least-squares line goes *up*.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

$-1 < r < 0$

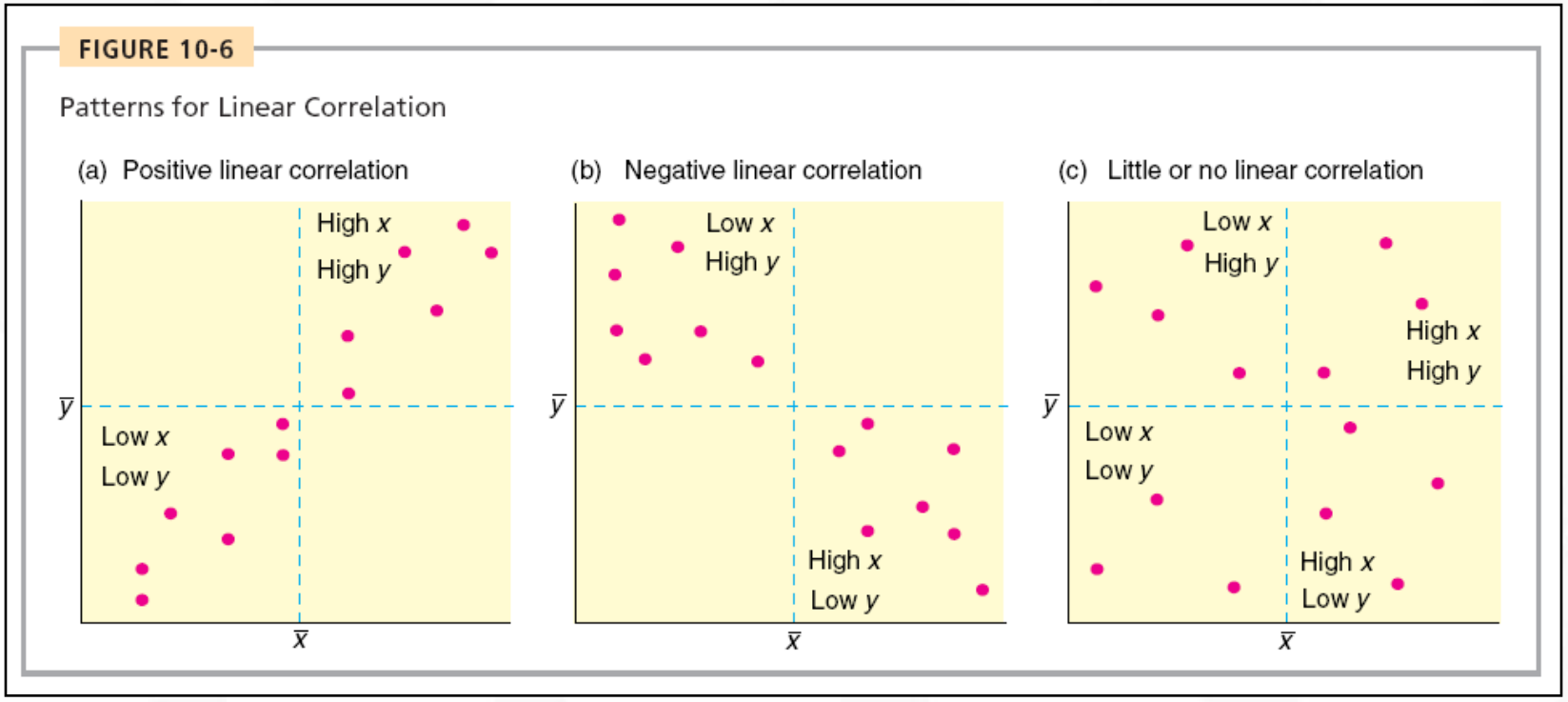
The x and y values have a *negative correlation.* By this, we mean *large* x values are associated with *small* y values, and *small* x values are associated with *large* y values.



As we go from left to right, the least-squares line goes *down.*

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Developing a Formula for *r*



Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Developing a Formula for *r*

$$r = \frac{1}{n-1} \sum \frac{(y - \bar{y})}{s_y} \cdot \frac{(x - \bar{x})}{s_x}$$

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

A Computational Formula for *r*

Obtain a random sample of n data pairs (x, y). The data pairs should have a *bivariate normal distribution*. This means that for a fixed value of x, the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed y, the x values should have their own (approximately) normal distribution.

1. Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy.

2. With n = sample size, Σx, Σy, Σx^2, Σy^2, and Σxy, you are ready to compute the sample correlation coefficient r using the computation formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the x values, then square the result.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Critical Thinking: Population vs. Sample Correlation

- Beware: high correlations **do not** imply a *cause and effect* relationship between the explanatory and response variables!!

r = **sample** correlation coefficient computed from a random sample of (x, y) data pairs.

ρ = **population** correlation coefficient computed from all population data pairs (x, y).

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Critical Thinking: Lurking Variables

- A lurking variable is neither the explanatory variable nor the response variable.
- Beware: lurking variables may be responsible for the evident changes in *x* and *y*.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Correlations Between Averages

- If the two variables of interest are averages rather than raw data values, the correlation between the averages will tend to be higher!

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Linear Regression

- When we are in the “paired-data” setting:
  - Do the data points have a linear relationship?
  - Can we find an equation for the best fitting line?
  - Can we predict the value of the response variable for a new value of the predictor variable?
  - What fractional part of the variability in *y* is associated with the variability in *x*?

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

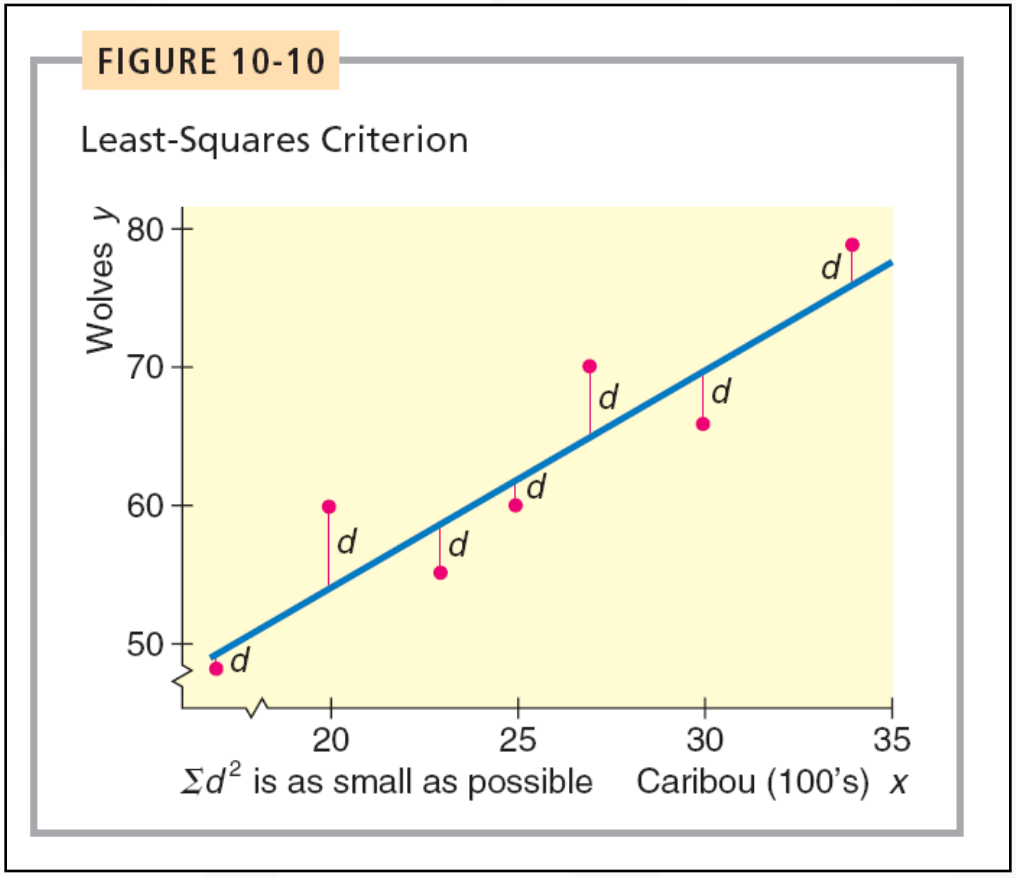
Least-Squares Criterion

- The line that we fit to the data will be such that the sum of the squared distances from the line will be minimized.

- Let *d* = the vertical distance from any data point to the regression line.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Least-Squares Criterion



Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Least-Squares Line

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

- *a* is the y-intercept.
- *b* is the slope.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Finding the Regression Equation

Obtain a random sample of n data pairs (x, y), where x is the *explanatory variable* and y is the *response variable*. The data pairs should have a *bivariate normal distribution*. This means that for a fixed value of x, the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed y, the x values should have their own (approximately) normal distribution.

1. Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy. Then compute the sample means $\overline{x}$ and $\overline{y}$.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Finding the Regression Equation

2. With n = sample size, Σx, Σy, Σx^2, Σy^2, Σxy, $\bar{x}$, and $\bar{y}$, you are ready to compute the slope b and intercept a using the computation formulas

Slope: $$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

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

Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the x values, then square the result.

3. The equation of the least-squares line computed from your sample data is

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

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Using the Regression Equation

- The point $(\bar{x}, \bar{y})$ is always on the least-squares line.

- Also, the slope tells us that if we increase the explanatory variable by one unit, the response variable will change by the slope (increase or decrease, depending on the sign of the slope).

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Critical Thinking: Making Predictions

- We can simply plug in *x* values into the regression equation to calculate *y* values.

Predicting $\hat{y}$ values for x values that are **between** observed x values in the data set is called **interpolation.**

Predicting $\hat{y}$ values of x values that are **beyond** observed x values in the data set is called **extrapolation.**

- Extrapolation can result in unreliable predictions.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Coefficient of Determination

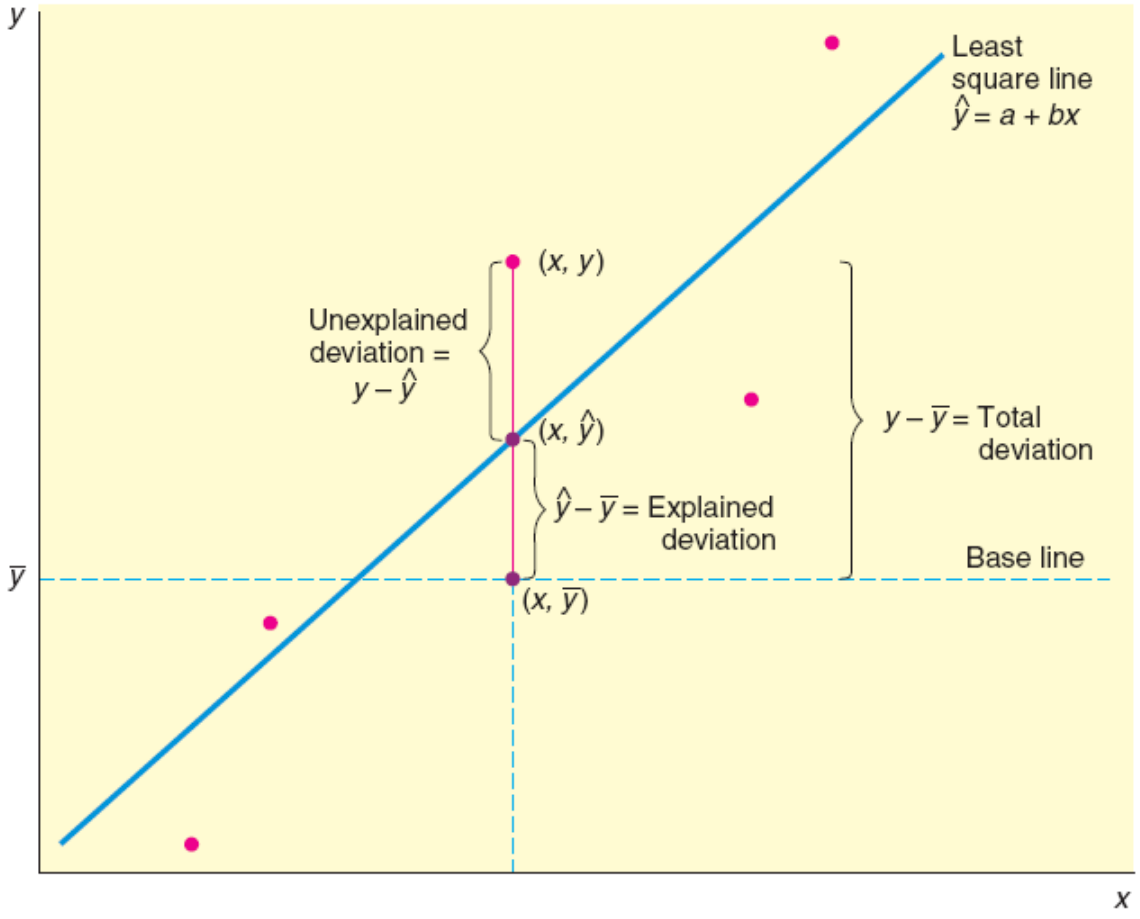
- Another way to gauge the fit of the regression equation is to calculate the coefficient of determination, r^2.
- We can view the *mean of y* as a baseline for all the *y* values.
- Then:

$$\text{Total deviation} = y - \bar{y}$$
$$\text{Explained deviation} = \hat{y} - \bar{y}$$
$$\text{Unexplained deviation} = y - \hat{y}$$

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

FIGURE 10-13

Explained and Unexplained Deviations



Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Coefficient of Determination

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

$$\begin{pmatrix}\text{Total}\\ \text{deviation}\end{pmatrix} = \begin{pmatrix}\text{Explained}\\ \text{deviation}\end{pmatrix} + \begin{pmatrix}\text{Unexplained}\\ \text{deviation}\end{pmatrix}$$

We algebraically manipulate the above equation to obtain:

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2$$

$$\begin{pmatrix}\text{Total}\\ \text{variation}\end{pmatrix} = \begin{pmatrix}\text{Explained}\\ \text{variation}\end{pmatrix} + \begin{pmatrix}\text{Unexplained}\\ \text{variation}\end{pmatrix}$$

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Coefficient of Determination

If r is the correlation coefficient [see Equation (2)], then it can be shown that

$$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} = \frac{\text{Explained variation}}{\text{Total variation}}$$

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

Thus, r^2 is the measure of the total variability in y that is explained by the regression on x.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Coefficient of Determination

Coefficient of determination r^2

1. Compute the sample correlation coefficient r using the procedure of Section 10.1. Then simply compute r^2, the sample coefficient of determination.
2. The value r^2 is the ratio of explained variation over total variation. That is, r^2 is the fractional amount of total variation in y that can be explained by using the linear model $\hat{y} = a + bx$.
3. Furthermore, $1 - r^2$ is the fractional amount of total variation in y that is due to random chance or to the possibility of lurking variables that influence y.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Inferences on the Regression Equation

- We can make inferences about the population correlation coefficient, ρ, and the population regression line slope, β.

Sample Statistic		Population Parameter
r	$\rightarrow$	ρ
a	$\rightarrow$	α
b	$\rightarrow$	β
$\hat{y} = a + bx$	$\rightarrow$	$y = \alpha + \beta x$

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Inferences on the Regression Equation

- Inference requirements:
  - The set of ordered pairs *(x, y)* constitutes a random sample from all ordered pairs in the population.
  - For each fixed value of *x*, *y* has a normal distribution.
  - All distributions for all *y* values have equal variance and a mean that lies on the regression equation.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Testing the Correlation Coefficient

H_0: x and y have no linear correlation, so $\rho = 0$

H_1: $\rho > 0$

H_1: $\rho < 0$

H_1: $\rho \neq 0$

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Testing the Correlation Coefficient

- We convert to a Student's t distribution.
- The sample size must be ≥ 3.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \text{with } d.f. = n - 2$$

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Testing Procedure

1. Use the *null hypothesis* H_0: $\rho = 0$. In the context of the application, state the *alternate hypothesis* ($\rho > 0$ or $\rho < 0$ or $\rho \neq 0$) and set the *level of significance* α.
2. Obtain a random sample of $n \geq 3$ data pairs (x, y) and compute the sample *test statistic*

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with degrees of freedom } d.f. = n - 2$$

3. Use a Student's t distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.
5. *State your conclusion* in the context of the application.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Standard Error of Estimate

$$\text{Standard error of estimate} = S_e = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}$$

where $\hat{y} = a + bx$ and $n \geq 3$.

- As the points get closer to the regression line, S_e decreases.
- If all the points lie exactly on the line, S_e will be zero.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Computational Formula For S_e

1. Obtain a random sample of $n \geq 3$ data pairs (x, y).
2. Use the procedures of Section 10.2 to find a and b from the sample least-squares line $\hat{y} = a + bx$.
3. The standard error of estimate is

$$S_e = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$$

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Confidence Intervals for *y*

- The population regression line:

  $y = \alpha + \beta x + \varepsilon$

  Where ε is random error
- Because of ε, for each x value there is a corresponding distribution for y.
- Each y distribution has the same standard deviation, estimated by S_e.
- Each y distribution is centered on the regression line.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Confidence Intervals for Predicted *y*

1. Obtain a random sample of $n \geq 3$ data pairs (x, y).
2. Use the procedure of Section 10.2 to find $\hat{y} = a + bx$. You also need to find $\overline{x}$ from the sample data and the standard error of estimate S_e using equation (8) of this section.
3. The c **confidence interval for** y for a **specified value of** x is

$$\hat{y} - E < y < \hat{y} + E$$

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Confidence Intervals for Predicted y

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n\Sigma x^2 - (\Sigma x)^2}}$$

$\hat{y}$ = $a + bx$ is the predicted value of y from the least-squares line for a *specified* x value

c = confidence level $(0 < c < 1)$

n = number of data pairs $(n \geq 3)$

t_c = critical value from Student's t distribution for c confidence level using $d.f. = n - 2$

S_e = standard error of estimate

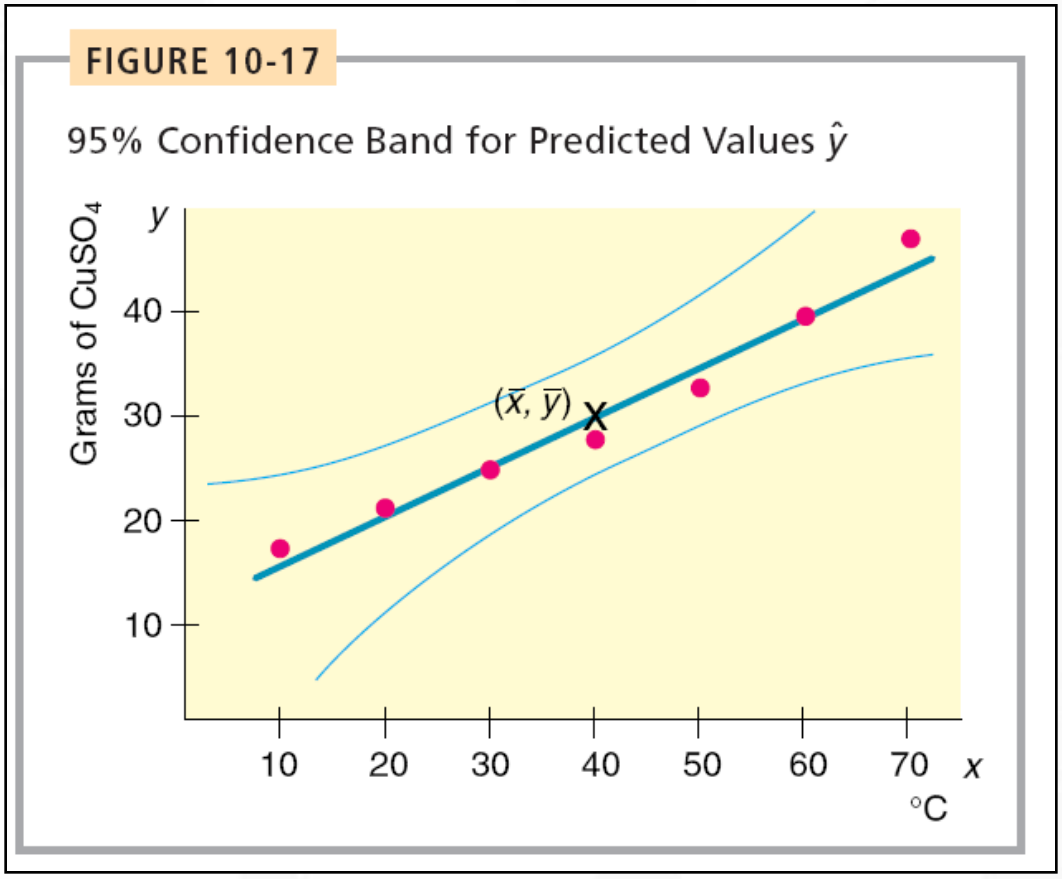
Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Confidence Intervals for Predicted *y*

- The confidence intervals will be narrower near the mean of *x*.
- The confidence intervals will be wider near the extreme values of *x*.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Confidence Intervals for Predicted *y*



Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Inferences About Slope β

- Recall the population regression equation: $y = \alpha + \beta x + \varepsilon$
- Let S_e be the standard error of estimate from the sample.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Inferences About Slope β

$$t = \frac{b - \beta}{S_e \Big/ \sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}}$$

- Has a Student's *t* distribution with *n* – *2* degrees of freedom.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Testing β

For a statistical test of $\boldsymbol{\beta}$

1. Use the *null hypothesis* H_0: $\beta = 0$. Use an *alternate hypothesis* H_1 appropriate to your application ($\beta > 0$ or $\beta < 0$ or $\beta \neq 0$). Set the level of significance α.
2. Use the null hypothesis H_0: $\beta = 0$ and the values of S_e, n, Σx, Σx^2, and b to compute the *sample test statistic*

$$t = \frac{b}{S_e}\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} \qquad \text{with } d.f. = n - 2$$

3. Use a Student's t distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If P-value $\leq \alpha$, then reject H_0. If P-value $> \alpha$, then do not reject H_0.
5. *State your conclusion* in the context of the application.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Finding a Confidence Interval for β

$$b - E < \beta < b + E$$

where $E = \dfrac{t_c S_e}{\sqrt{\Sigma x^2 - \dfrac{1}{n}(\Sigma x)^2}}$

c = confidence level $(0 < c < 1)$

n = number of data pairs (x, y), $n \geq 3$

t_c = Student's t distribution critical value for confidence level c and $d.f. = n - 2$

S_e = standard error of estimate

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Multiple Regression

- In practice, most predictions will improve if we consider additional data.
- Statistically, this means adding predictor variables to our regression techniques.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Multiple Regression Terminology

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

- y is the response variable.
- $x_1 \ldots x_k$ are explanatory variables.
- $b_0 \ldots b_k$ are coefficients.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Regression Models –

A mathematical package that includes the following:

1) A collection of variables, one of which is the response, and the rest are predictors.
2) A collection of numerical values associated with the given variables.
3) Using software and the above numerical values, construct a *least-squares regression equation*.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Regression Models –

A mathematical package that includes the following:

4) The model includes information about the variables, coefficients of the equation, and various “goodness of fit” measures.
5) The model allows the user to make predictions and build confidence intervals for various components of the equation.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Coefficient of Multiple Determination

- How well does our equation fit the data?
  - Just as in linear regression, computer output will provide r^2 .
    - r^2 is a measure of the amount of variability in *y* that is explained by the regression on all of the *x* variables.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Predictions in the Multiple Regression Case

- Similar to the techniques described for linear regression, we can “plug in” values for all of the *x* variables and compute a predicted *y* value.

  – Predicting *y* for any *x* values outside the *x*-range is extrapolation, and the results will be unreliable.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Testing a Coefficient for Significance

- At times, one or more of the predictor variables may not be helpful in determining *y*.
- Using software, we can test any or all of the *x* variables for statistical significance.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Testing a Coefficient for Significance

- The hypotheses will be:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

for the i^{th} coefficient in the model.

- If we fail to reject the null hypothesis for a given coefficient, we may consider removing that predictor from the model.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.

Confidence Intervals for Coefficients

- As in the linear case, software can provide confidence intervals for any β_i in the model.

Copyright © Houghton Mifflin Harcourt Publishing Company. All rights reserved.