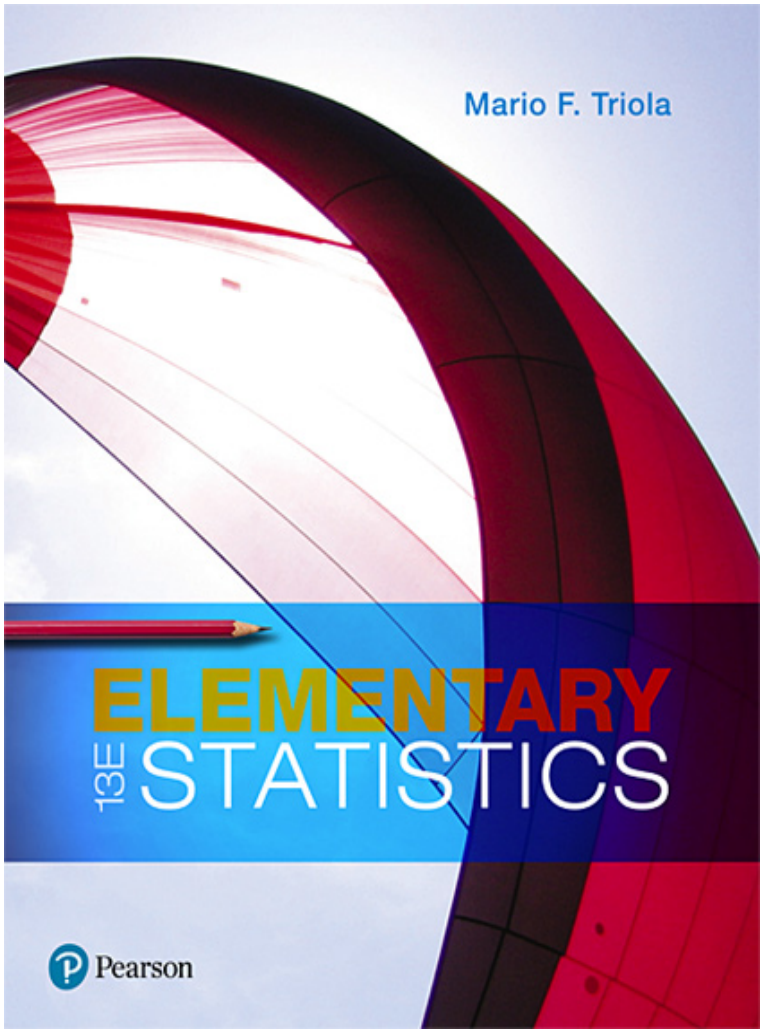
Elementary Statistics

Thirteenth Edition



Chapter 10

Correlation and Regression

Correlation and Regression

10-1 Correlation

10-2 Regression

10-3 Prediction Intervals and Variation

10-4 Multiple Regression

10-5 Nonlinear Regression

Key Concepts

This section presents methods for finding the equation of the straight line that best fits the points in a scatterplot of paired sample data. That best-fitting straight line is called the regression line, and its equation is called the regression equation. Then we discuss marginal change, influential points, and residual plots as tools for analyzing correlation and regression results.

Regression Line

- Regression Line
  - Given a collection of paired sample data, the **regression line** (or line of best fit, or least-squares line) is the straight line that “best” fits the scatterplot of the data.

Regression Equation

- Regression Equation
  - The **regression equation**

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

algebraically describes the regression line. The regression equation expresses a relationship between x (called the **explanatory variable,** or **predictor variable,** or **independent variable**) and $\hat{y}$ (called the **response variable** or **dependent variable**).

Finding the Equation of the Regression Line: Objective

Find the equation of a regression line.

Finding the Equation of the Regression Line: Notation for the Equation of a Regression Line

	Sample Statistic	Population Parameter
y-intercept of regression equation	b_0	β_0
Slope of regression equation	b_1	β_1
Equation of the regression line	$\hat{y} = b_0 + b_1 x$	$y = \beta_0 + \beta_1 x$

Finding the Equation of the Regression Line: Requirements (1 of 2)

1. The sample of paired (x, y) data is a **random** sample of quantitative data.
2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.*
3. Outliers can have a strong effect on the regression equation, so remove any outliers if they are known to be errors. Consider the effects of any outliers that are not known errors.*

Finding the Equation of the Regression Line: Requirements (2 of 2)

*Note: Requirements 2 and 3 are simplified attempts at checking these formal requirement for regression analysis:

- For each fixed value of *x*, the corresponding values of y have a normal distribution.
- For the different fixed values of *x*, the distributions of the corresponding *y*-values all have the same standard deviation. (This is violated if part of the scatterplot shows points very close to the regression line while another portion of the scatterplot shows points that are much farther away from the regression line.)
- For the different fixed values of *x*, the distributions of the corresponding *y* values have means that lie along the same straight line.

Finding the Equation of the Regression Line: Formulas for Finding the Slope b_1 and y-Intercept b_0 in the Regression Equation y-hat = $b_0 + b_1x$ (1 of 2)

Slope: $b_1 = r\dfrac{s_y}{s_x}$

y-intercept: $b_0 = \bar{y} - b_1\bar{x}$

Finding the Equation of the Regression Line: Formulas for Finding the Slope b_1 and y-Intercept b_0 in the Regression Equation y-hat = $b_0 + b_1x$ (2 of 2)

The slope b_1 and y-intercept b_0 can also be found using the following formulas that are useful for manual calculations or writing computer programs:

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \qquad b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

Finding the Equation of the Regression Line: Rounding the Slope b_1 and the y-Intercept b_0

Round b_1 and b_0 to three significant digits. It's difficult to provide a simple universal rule for rounding values of b_1 and b_0, but this rule will work for most situations.

Example: Using Technology to Find the Regression Equation (1 of 11)

Use technology to find the equation of the regression line in which the explanatory variable (or *x* variable) is chocolate consumption and the response variable (or *y* variable) is the corresponding Nobel Laureate rate. The table of data is on the next slide.

Example: Using Technology to Find the Regression Equation (2 of 11)

Chocolate	Nobel
4.5	5.5
10.2	24.3
4.4	8.6
2.9	0.1
3.9	6.1
0.7	0.1
8.5	25.3
7.3	7.6
6.3	9.0
11.6	12.7
2.5	1.9
8.8	12.7

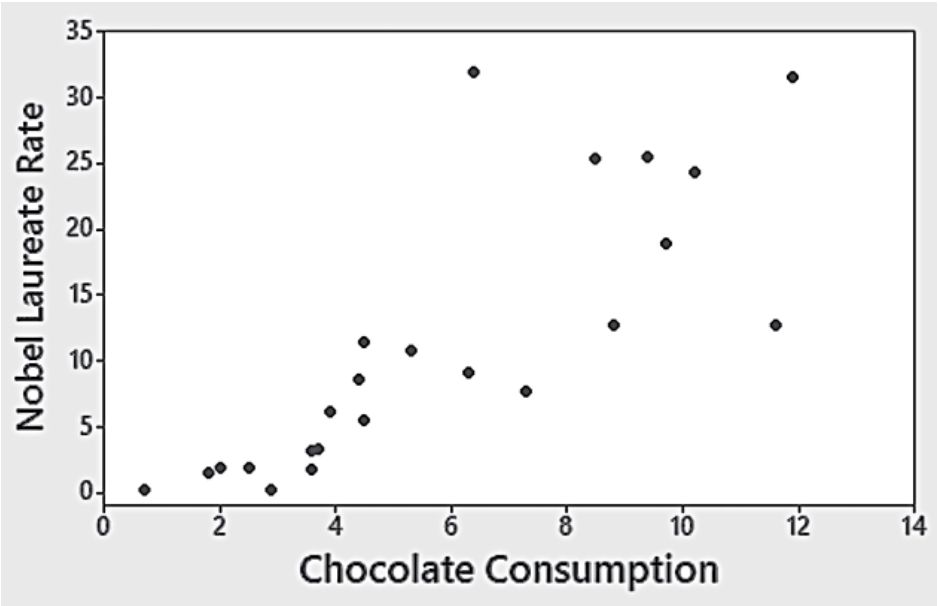
Example: Using Technology to Find the Regression Equation (3 of 11)

Chocolate	Nobel
3.7	3.3
1.8	1.5
4.5	11.4
9.4	25.5
3.6	3.1
2.0	1.9
3.6	1.7
6.4	31.9
11.9	31.5
9.7	18.9
5.3	10.8

Example: Using Technology to Find the Regression Equation (4 of 11)

Solution

REQUIREMENT CHECK
(1) The data are assumed to be a simple random sample. (2) The figure is a scatterplot showing a pattern of points. This pattern is very roughly a straight-line pattern. (3) There are no outliers. The requirements are satisfied.



Example: Using Technology to Find the Regression Equation (5 of 11)

Solution

Technology The use of technology is recommended for finding the equation of a regression line. On the next slide are the results from different technologies. Minitab and XLSTAT provide the actual equation; the other technologies list the values of the *y*-intercept and the slope. All of these technologies show that the regression equation can be expressed as $\hat{y} = -3.37 + 2.49x$, where $\hat{y}$ is the predicted Nobel Laureate rate and x is the amount of chocolate consumption.

Example: Using Technology to Find the Regression Equation (6 of 11)

Solution

Statdisk

Regression Results:
Y= b0 + b1x:
Y Intercept, b0: −3.366668
Slope, b1: 2.493134

Excel (XLSTAT)

Equation of the model:

Nobel = -3.36667+2.49313*Choc

Minitab

Regression Equation

Nobel = −3.37 + 2.493 Choc

TI-83/84 Plus

NORMAL FLOAT AUTO REAL RADIAN MP

LinRegTTest
y=a+bx
β≠0 and ρ≠0
t=6.123022389
p=4.4774676E-6
df=21
a=-3.366667586
b=2.493133741
↓s=6.262664763

SPSS

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-3.367	2.700		-1.247	.226
	Choc	2.493	.407	.801	6.123	.000

JMP

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-3.366668	2.700151	-1.25	0.2262
Choc	2.4931337	0.407174	6.12	<.0001*

StatCrunch

Simple linear regression results:
Dependent Variable: Nobel
Independent Variable: Choc
Nobel = -3.3666676 + 2.4931337 Choc

Example: Using Technology to Find the Regression Equation (7 of 11)

Solution

We should know that the regression equation is an **estimate** of the true regression equation for the population of paired data. This estimate is based on one particular set of sample data, but another sample drawn from the same population would probably lead to a slightly different equation.

Example: Using Technology to Find the Regression Equation (8 of 11)

Use the first formulas for b_1 and b_0 to find the equation of the regression line in which the explanatory variable (or x variable) is chocolate consumption and the response variable (or y variable) is the corresponding number of Nobel Laureates.

Example: Using Technology to Find the Regression Equation (9 of 11)

Solution

REQUIREMENT CHECK The requirements are previously verified.

We begin by finding the slope b_1 as follows. Remember, r is the linear correlation coefficient, s_y is the standard deviation of the sample y values, and s_x is the standard deviation of the sample x values.

$$b_1 = r\frac{s_y}{s_x} = 0.800608 \cdot \frac{10.211601}{3.279201} = 2.493135$$

Example: Using Technology to Find the Regression Equation (10 of 11)

Solution

After finding the slope b_1, we can now find the y-intercept as follows:

$$b_0 = \bar{y} - b_1\bar{x}$$
$$= 11.104348 - (2.493135)(5.804348)$$
$$= -3.366675$$

Example: Using Technology to Find the Regression Equation (11 of 11)

Solution

After rounding, the slope is $b_1 = 2.49$ and the y-intercept is $b_0 = -3.37$. We can now express the regression equation as $\hat{y} = -3.37 + 2.49x$, where $\hat{y}$ is the predicted Nobel Laureate rate and x is the amount of chocolate consumption.

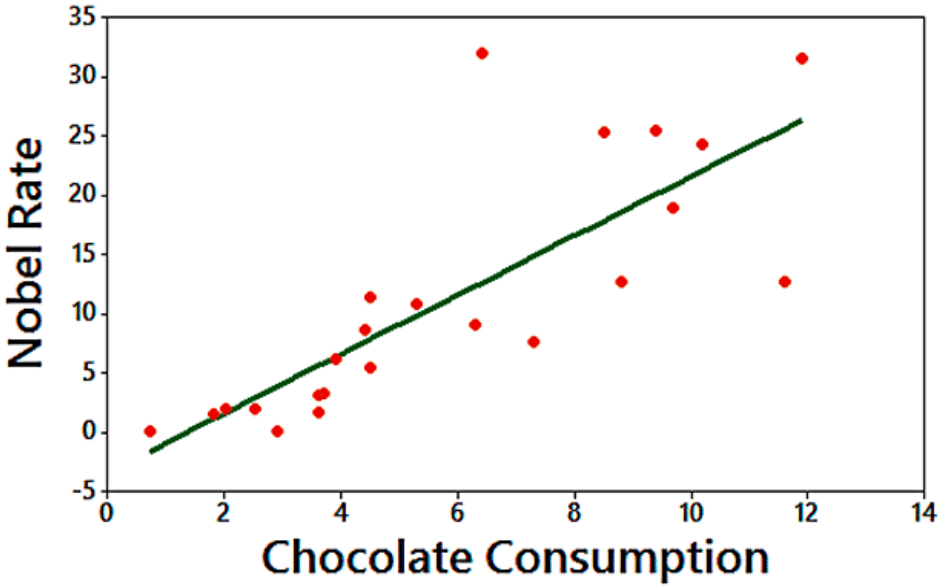
Example: Graphing the Regression Line (1 of 2)

Graph the regression equation $\hat{y} = -3.37 + 2.49x$ on the scatterplot of the chocolate/Nobel data and examine the graph to subjectively determine how well the regression line fits the data.

Example: Graphing the Regression Line (2 of 2)

Solution

Shown below is the Minitab display of the scatterplot with the graph of the regression line included. We can see that the regression line fits the points well, but the points are not very close to the line.



Making Predictions (1 of 2)

Regression equations are often useful for predicting the value of one variable, given some specific value of the other variable. When making predictions, we should consider the following:

1. **Bad Model:** If the regression equation does not appear to be useful for making predictions, don't use the regression equation for making predictions. For bad models, the best predicted value of a variable is simply its sample mean.
2. **Good Model:** Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.

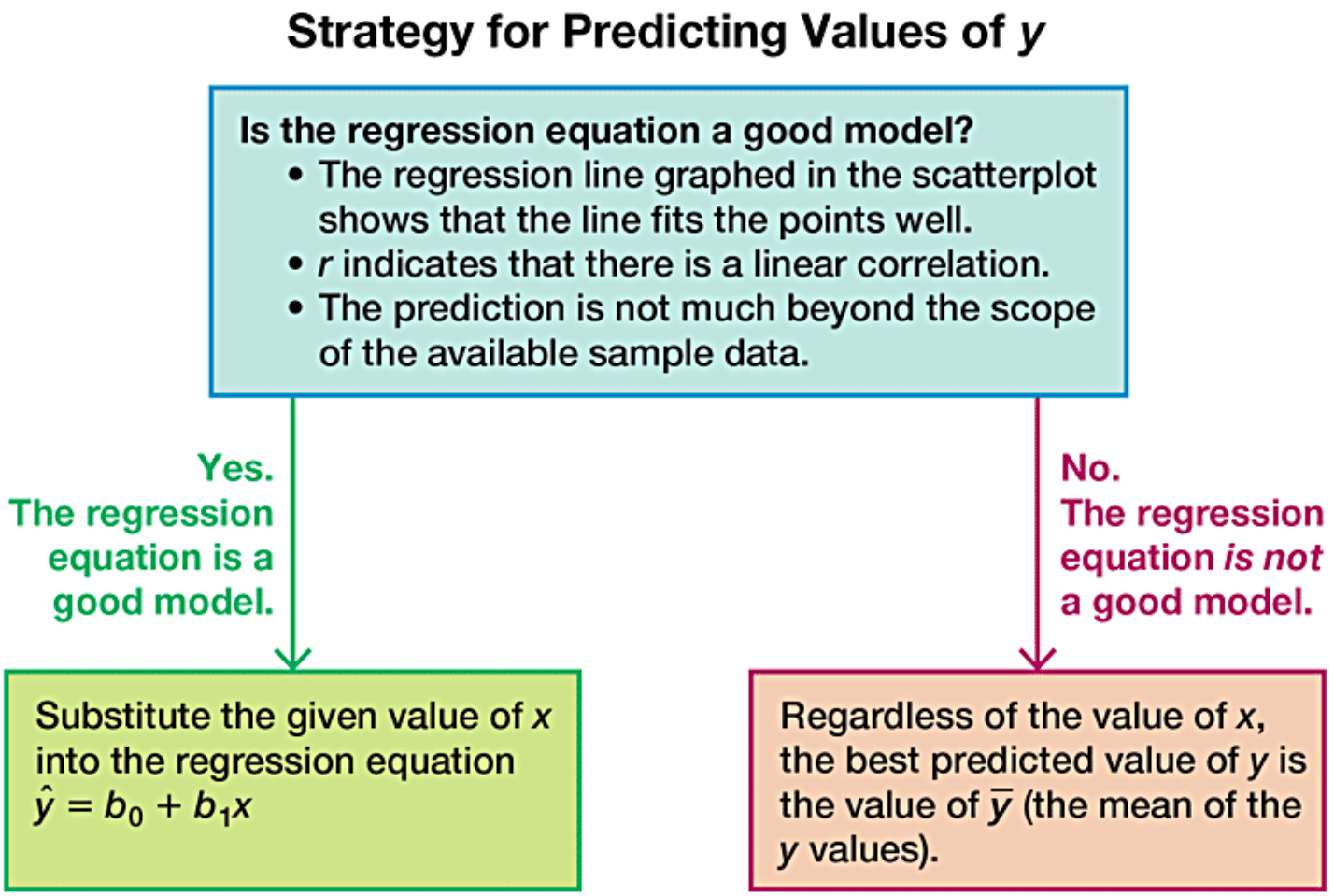
Making Predictions (2 of 2)

3. **Correlation:** Use the regression equation for predictions only if the linear correlation coefficient *r* indicates that there is a linear correlation between the two variables.

4. **Scope:** Use the regression line for predictions only if the data do not go much beyond the scope of the available sample data.

The figure on the next slide summarizes a strategy for predicting values of a variable *y* when given some value of *x*.

Strategy for Prediction Values of *y*



Example: Making Predictions (1 of 4)

a. Use the chocolate/Nobel data to predict the Nobel rate for a country with chocolate consumption of 10 kg per capita.

b. Predict the IQ score of an adult who is exactly 175 cm tall.

Example: Making Predictions (2 of 4)

Solution

a. **Good Model: Use the Regression Equation for Predictions.**

The regression line fits the points well. Also, there is a linear correlation between chocolate consumption and the Nobel Laureate rate. Because the regression equation $\hat{y} = -3.37 + 2.49x$ is a good model, substitute $x = 10$ into the regression equation to get a predicted Nobel Laureate rate of 21.5 Nobel Laureates per 10 million people.

Example: Making Predictions (3 of 4)

Solution

b. **Bad Model: Use *y* for predictions.** Knowing that there is no correlation between height and IQ score, we know that a regression equation is not a good model, so the best predicted value of IQ score is the mean, which is 100.

Example: Making Predictions (4 of 4)

Interpretation

In part (a), the paired data result in a **good** regression model, so the predicted Nobel rate is found by substituting the value of $x = 10$ into the regression equation.

In part (b) there is no correlation between height and IQ, so the best predicted IQ score is the mean of $\bar{y} = 100$.

Key point: Use the regression equation for predictions only if it is a good model. If the regression equation is not a good model, use the predicted value of $\bar{y}$.

Marginal Change (1 of 2)

- Marginal Change
  - In working with two variables related by a regression equation, the **marginal change** in a variable is the amount that it changes when the other variable changes by exactly one unit. The slope b_1 in the regression equation represents the marginal change in y that occurs when x changes by one unit.

Marginal Change (2 of 2)

Let's consider the 23 pairs of chocolate/Nobel data. Those 23 pairs of data result in this regression equation: $\hat{y} = -3.37 + 2.49x$.

The slope of 2.49 tells us that if we increase x (chocolate consumption) by 1 (kg per capita), the predicted Nobel Laureate rate will increase by 2.49 (per 10 million people).

That is, for every additional 1 kg per capita increase in chocolate consumption, we expect the Nobel Laureate rate to increase by 2.49 per 10 million people.

Outlier

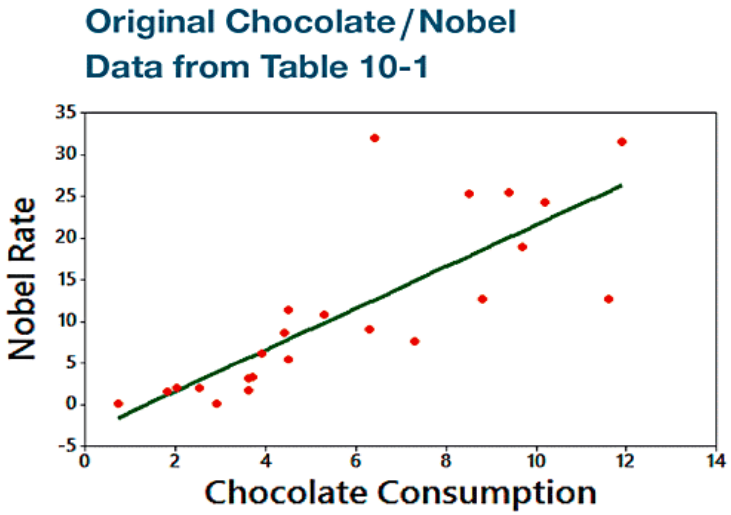
- **Outlier**
  - In a scatterplot, an **outlier** is a point lying far away from the other data points.

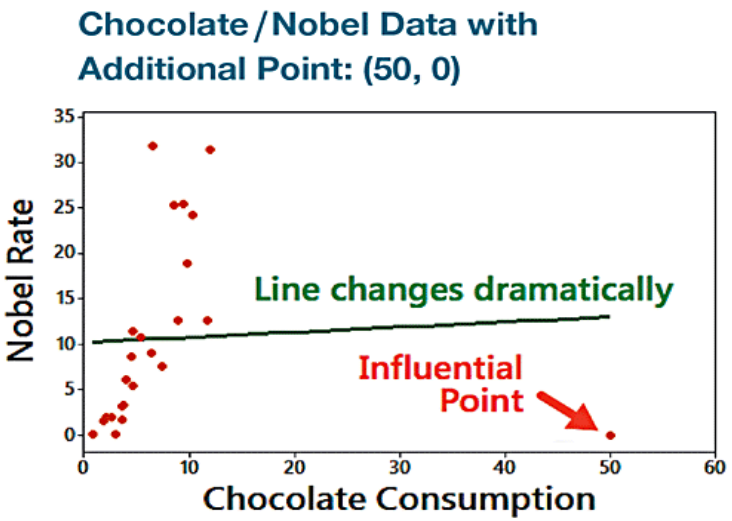
Influential Points

- **Influential Points**
  - Paired sample data may include one or more **influential points,** which are points that strongly affect the graph of the regression line.

Example: Influential Points (1 of 2)

Consider the 23 pairs of chocolate/Nobel data. The scatterplot located to the left below shows the regression line. If we include an additional pair of data, $x = 50$ and $y = 0$, we get the regression line shown to the right below.





Pearson

Example: Influential Points (2 of 2)

The additional point (50,0) is an influential point because the graph of the regression line did change considerably, as shown by the regression line on the previous screen. Compare the two graphs to see clearly that the addition of this one pair of values has a very dramatic effect on the regression line, so that additional point is an influential point. The additional point is also an outlier because it is far from the other points.

Residual

- Residual
  - For a pair of sample x and y values, the **residual** is the difference between the **observed** sample value of y and the y value that is **predicted** by using the regression equation.

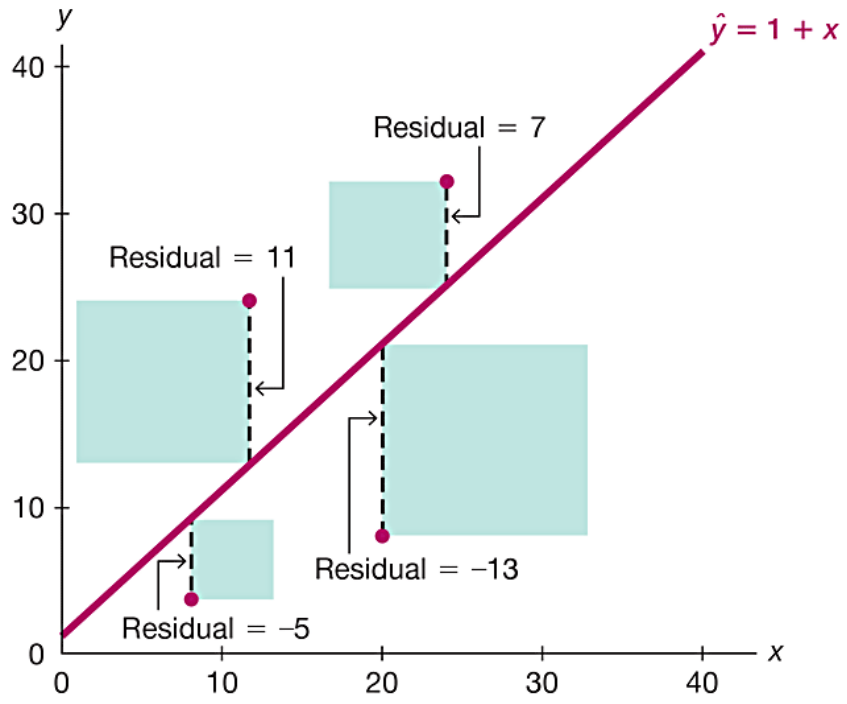
    That is,

    Residual = observed y − predicted y = $y - \hat{y}$

Residuals

Consider the sample point with coordinates of (8, 4). If we substitute $x = 8$ into the regression equation $\hat{y} = 1 + x$, we get a predicted value of $\hat{y} = 9$. But for $x = 8$, the actual observed sample value is $y = 4$. The difference $y - \hat{y} = 4 - 9 = -5$ is a residual.

x	8	12	20	24
y	4	24	8	32



Least-Squares Property

- Least-Squares Property
  - A straight line satisfies the **least-squares property** if the sum of the squares of the residuals is the smallest sum possible.

Residual Plots (1 of 2)

- Residual Plots
  - A **residual plot** is a scatterplot of the (x, y) values after each of the y-coordinate values has been replaced by the residual value $y - \hat{y}$ (where $\hat{y}$ denotes the predicted value of y). That is, a residual plot is a graph of the points $(x, y - \hat{y})$.

Residual Plots (2 of 2)

To construct a residual plot, draw a horizontal reference line through the residual value of 0, then plot the paired values of $(x, y - \hat{y})$.

Analyzing a Residual Plot

- The residual plot should not have any obvious pattern
- The residual plot should not become much wider (or thinner) when viewed from left to right.

Example: Residual Plots

The chocolate/Nobel data are used to obtain the Statdisk-generated residual plot on the next slide. When the first sample x value of 4.5 is substituted into the regression equation of $\hat{y} = -3.37 + 2.49x$, we get the predicted value of $\hat{y} = 7.84$. For the first x value of 4.5, the actual corresponding y value is 5.5, so the value of the residual is

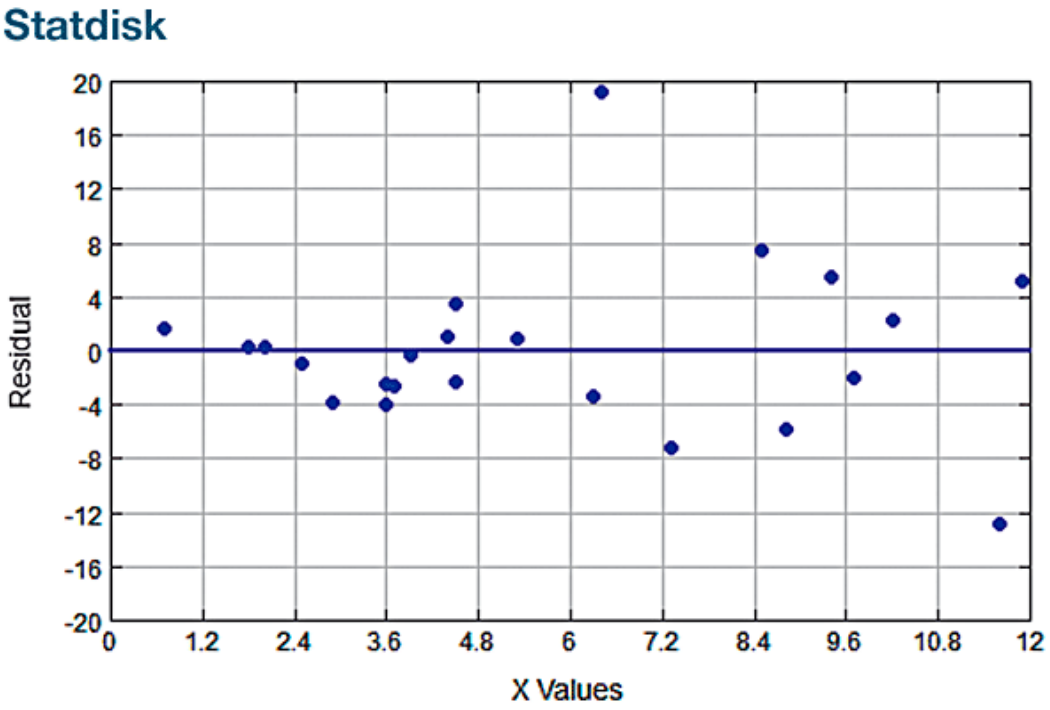
$$\text{Observed } y - \text{predicted } y = y - \hat{y}$$

$$= 5.5 - 7.84$$

$$= -2.34$$

Example: Residual Plots (2 of 2)

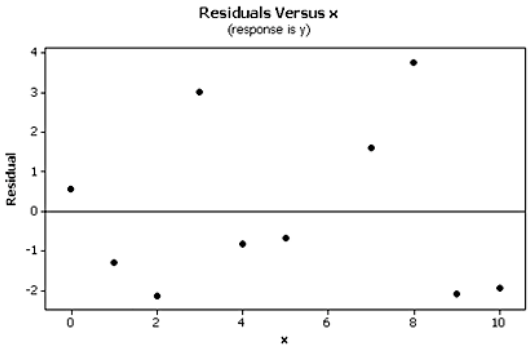
Using the *x* value of 4.5 and the residual of −2.34, we get the coordinates of the point (4.5, −2.34), which is one of the points in the residual plot shown here.



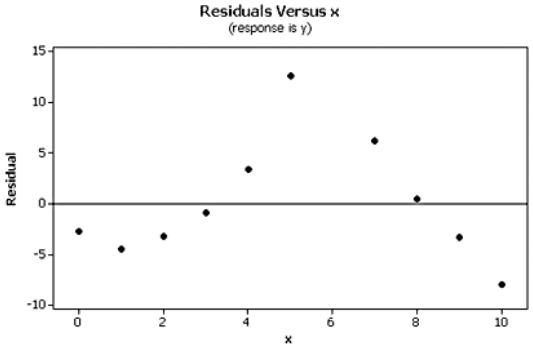
Residual Plots (3 of 3)

See the three residual plots below. The leftmost residual plot suggests that the regression equation is a good model. The middle residual plot shows a distinct pattern, suggesting that the sample data do not follow a straight-line pattern as required. The rightmost residual plot becomes thicker, which suggests that the requirement of equal standard deviations is violated.

Residual Plot Suggesting That the Regression Equation Is a Good Model



Residual Plot with an Obvious Pattern, Suggesting That the Regression Equation Is Not a Good Model



Residual Plot That Becomes Wider, Suggesting That the Regression Equation Is Not a Good Model