

AMA 2634

Introduction to Statistics

2024 Fall

Lecture 10

Regression

In addition to hypothesis testing and confidence intervals, inferential statistics involves determining whether a relationship between two or more numerical or quantitative variables exists.

Correlation is a statistical method used to determine whether a linear relationship between variables exists.

Regression is a statistical method used to describe the nature of the relationship between variables—that is, positive or negative, linear or nonlinear.

The purpose of this chapter is to answer these questions statistically:

1. Are two or more variables related?
2. If so, what is the strength of the relationship?
3. What type of relationship exists?
4. What kind of predictions can be made from the relationship?

- 1. Are two or more variables related?*
- 2. If so, what is the strength of the relationship?*

To answer these two questions, statisticians use the **correlation coefficient**, a numerical measure to determine whether two or more variables are related and to determine the strength of the relationship between or among the variables.

3. What type of relationship exists?

There are two types of relationships: simple and multiple.

In a simple relationship, there are two variables: an **independent variable** (predictor variable) and a **dependent variable** (response variable).

In a multiple relationship, there are two or more independent variables that are used to predict one dependent variable.

4. What kind of predictions can be made from the relationship?

Predictions are made daily in all areas. Examples include weather forecasting, stock market analyses, sales predictions, crop predictions, gasoline price predictions, and sports predictions. Some predictions are more accurate than others, due to the strength of the relationship. That is, the stronger the relationship is between variables, the more accurate the prediction is.



A **scatter plot** is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y .

Example: Car Rental Companies



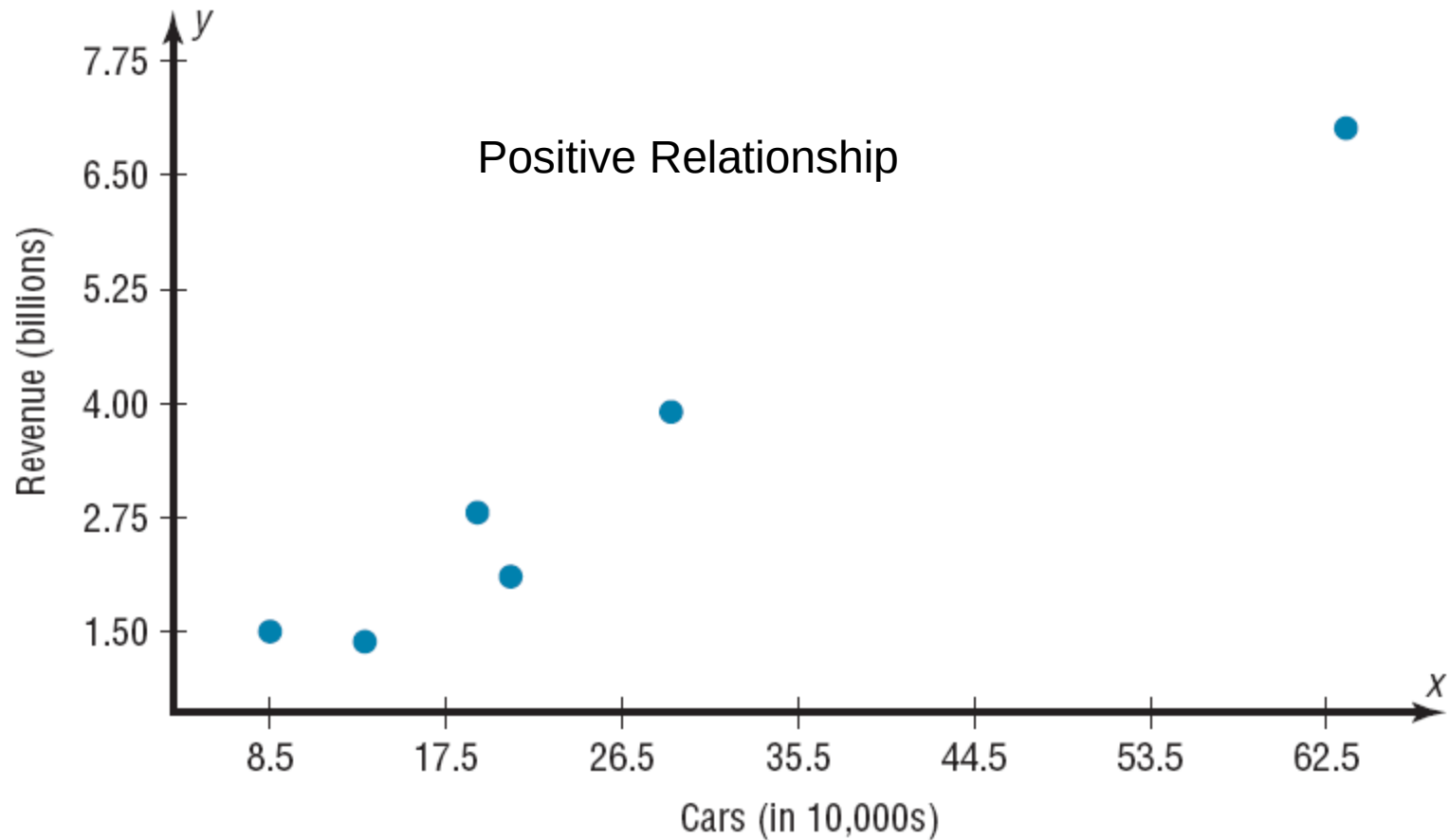
Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	\$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Step 1: Draw and label the x and y axes.

Step 2: Plot each point on the graph.

Example: Car Rental Companies



Example: Absences/Final Grades



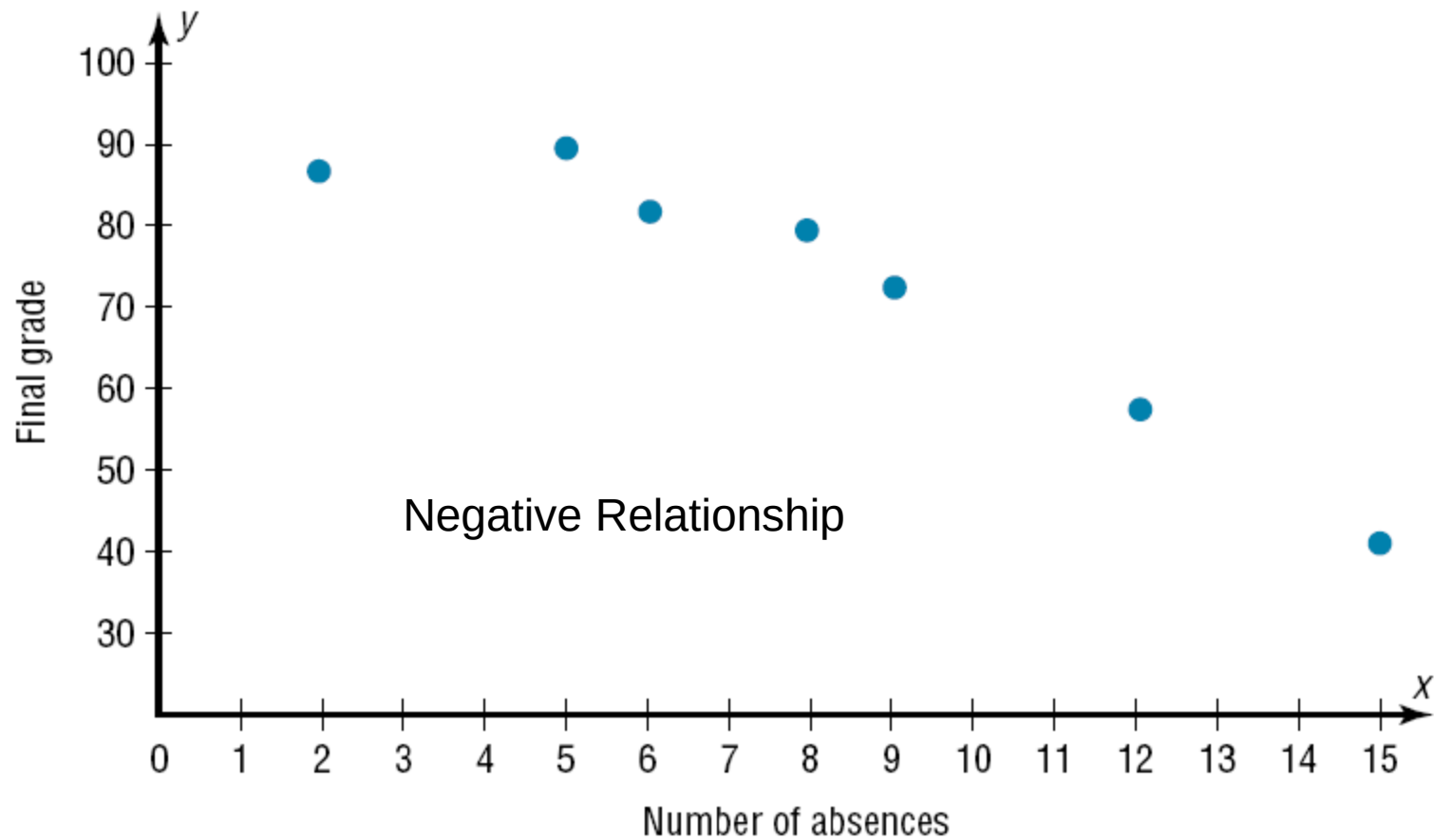
Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

Student	Number of absences x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

Step 1: Draw and label the x and y axes.

Step 2: Plot each point on the graph.

Example: Absences/Final Grades



Example: Age and Wealth



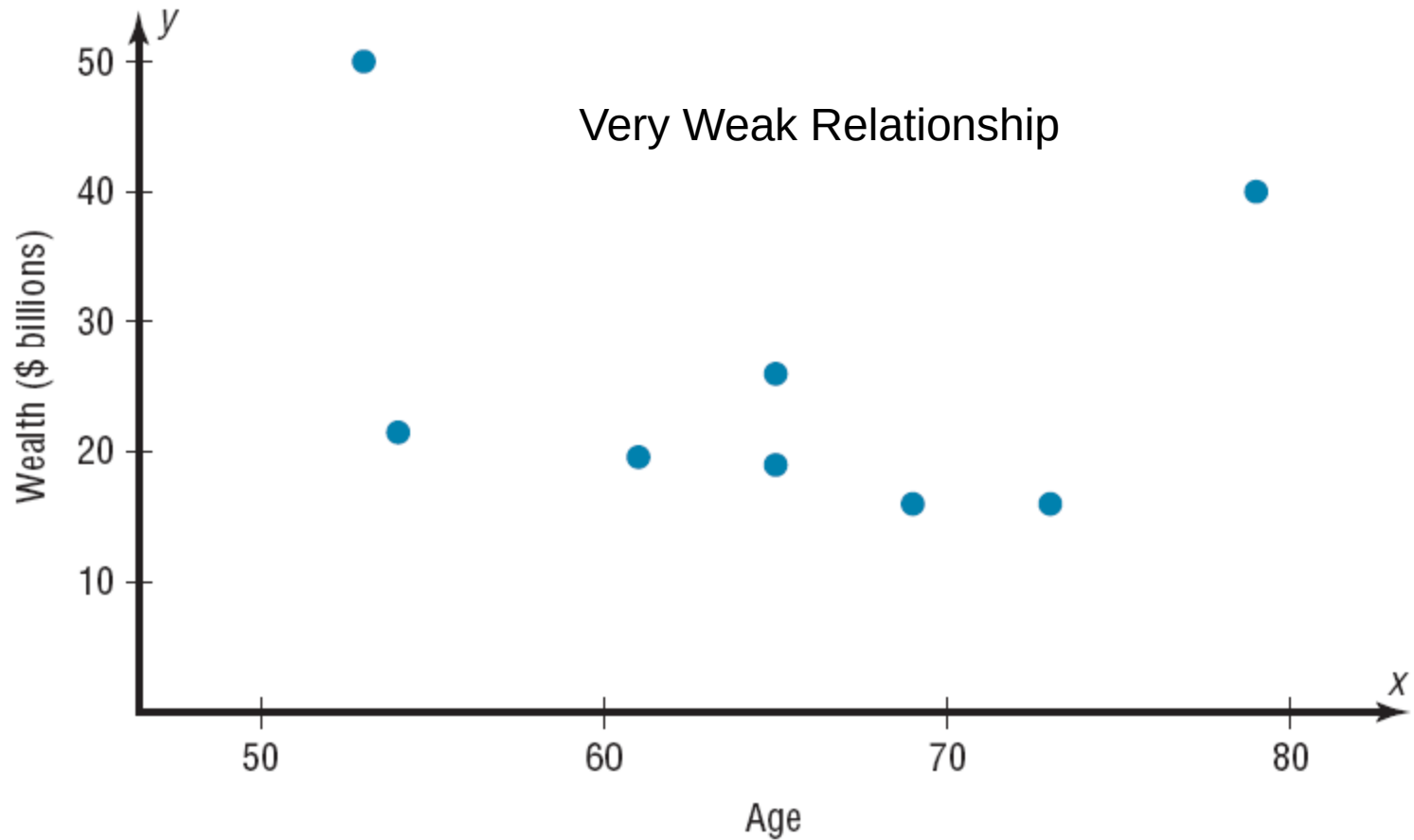
A researcher wishes to see if there is a relationship between the ages and net worth of the wealthiest people in America. The data for a specific year are shown.

Person	Age x	Net wealth y (\$ billions)
A	73	16
B	65	26
C	53	50
D	54	21.5
E	79	40
F	69	16
G	61	19.6
H	65	19

Step 1: Draw and label the x and y axes.

Step 2: Plot each point on the graph.

Example: Age and Wealth



The **correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two variables.

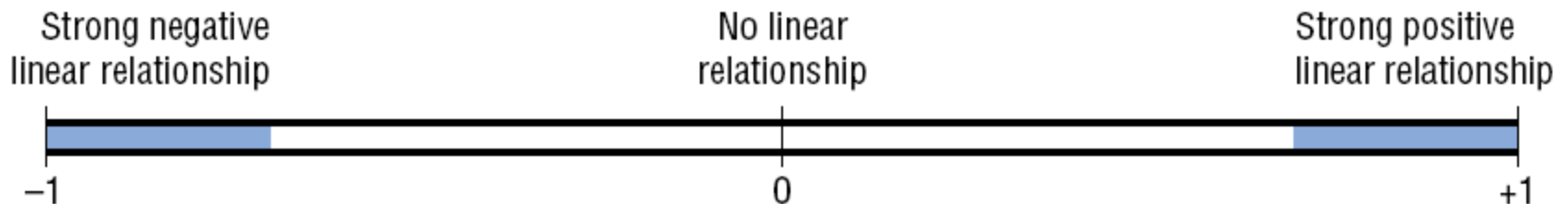
There are several types of correlation coefficients. The one explained in this section is called the **Pearson product moment correlation coefficient (PPMC)**.

The symbol for the sample correlation coefficient is r .
The symbol for the population correlation coefficient is ρ .

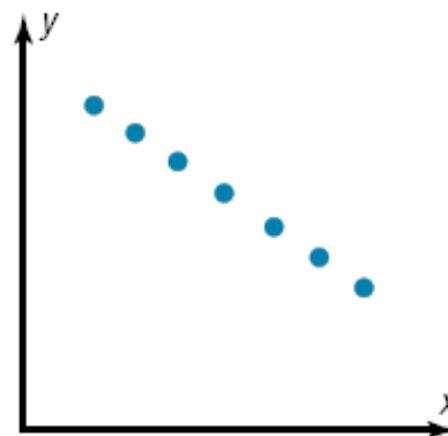
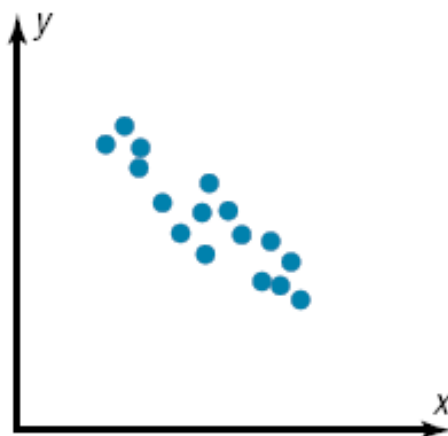
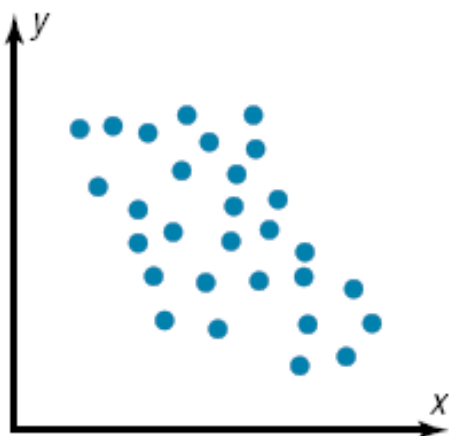
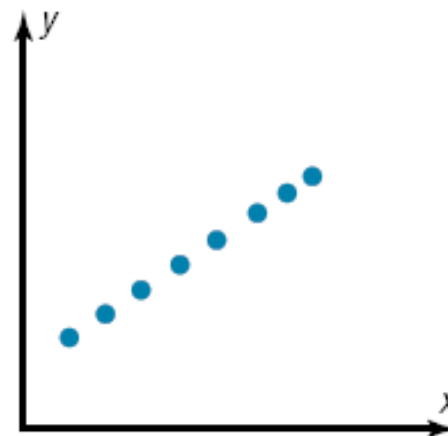
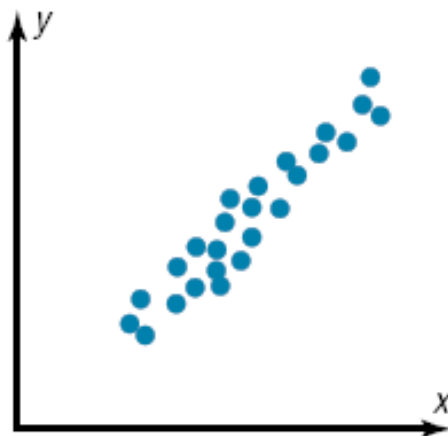
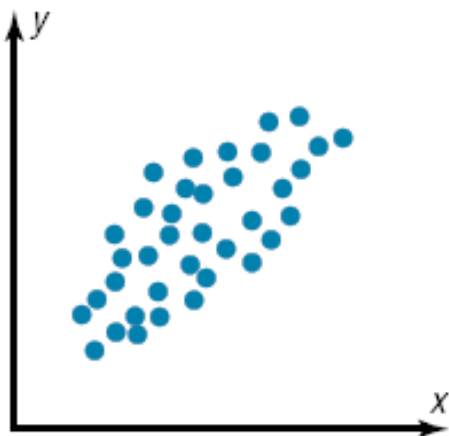
The range of the correlation coefficient is from -1 to $+1$.

If there is a **strong positive linear relationship** between the variables, the value of r will be close to $+1$.

If there is a **strong negative linear relationship** between the variables, the value of r will be close to -1 .



Correlation



The formula for the correlation coefficient is

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n(\sum x^2) - (\sum x)^2\right]\left[n(\sum y^2) - (\sum y)^2\right]}}$$

where n is the number of data pairs.

Can use the function Correl in Excel to compute r !

Example: Car Rental Companies



Company	Cars x (in 10,000s)	Income y (in billions)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	12.75	72.25	2.25
	$\Sigma x =$ 153.8	$\Sigma y =$ 18.7	$\Sigma xy =$ 682.77	$\Sigma x^2 =$ 5859.26	$\Sigma y^2 =$ 80.67

$$\Sigma x = 153.8, \Sigma y = 18.7, \Sigma xy = 682.77, \Sigma x^2 = 5859.26, \\ \Sigma y^2 = 80.67, n = 6$$

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

$$r = \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}}$$

$$r = 0.982 \text{ (strong positive relationship)}$$

In hypothesis testing, one of the following is true:

$H_0: \rho = 0$ This null hypothesis means that there is no correlation between the x and y variables in the population.

$H_1: \rho \neq 0$ This alternative hypothesis means that there is a significant correlation between the variables in the population.



$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

with degrees of freedom equal to $n - 2$.



Test the significance of the correlation coefficient found in the car rental example. Use $\alpha = 0.05$ and $r = 0.982$.

Step 1: State the hypotheses.

$$H_0: \rho = 0 \text{ and } H_1: \rho \neq 0$$

Step 2: Find the critical value.

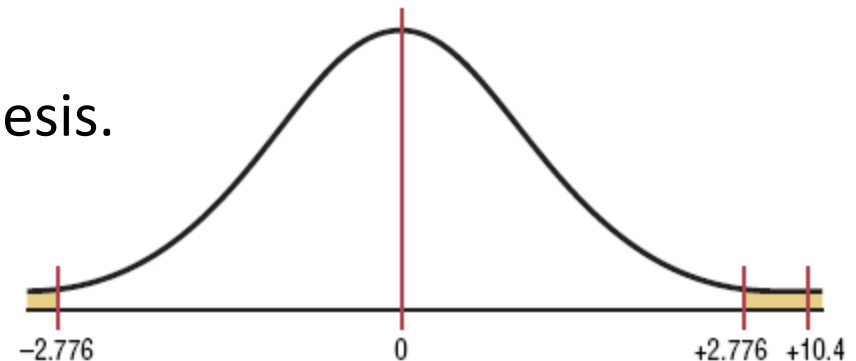
Since $\alpha = 0.05$ and there are $6 - 2 = 4$ degrees of freedom, the critical values obtained from Table F are ± 2.776 .

Step 3: Compute the test value.

$$t = r \sqrt{\frac{n-2}{1-r^2}} = 0.982 \sqrt{\frac{6-2}{1-(0.982)^2}} = 10.4$$

Step 4: Make the decision.

Reject the null hypothesis.



Step 5: Summarize the results.

There is a significant relationship between the number of cars a rental agency owns and its annual income.

When the null hypothesis has been rejected for a specific a value, any of the following five possibilities can exist.

1. There is a *direct cause-and-effect* relationship between the variables. That is, x causes y .
2. There is a *reverse cause-and-effect* relationship between the variables. That is, y causes x .
3. The relationship between the variables may be *caused by a third variable*.
4. There may be a *complexity of interrelationships* among many variables.
5. The relationship may be *coincidental*.



1. There is a *direct cause-and-effect* relationship between the variables. That is, x causes y .

For example,

- water causes plants to grow
- poison causes death
- heat causes ice to melt



2. There is a *reverse cause-and-effect* relationship between the variables. That is, y causes x .

For example,

- Suppose a researcher believes excessive coffee consumption causes nervousness, but the researcher fails to consider that the reverse situation may occur. That is, it may be that an extremely nervous person craves coffee to calm his or her nerves.

3. The relationship between the variables may be *caused by a third variable*.

For example,

- If a statistician correlated the number of deaths due to drowning and the number of cans of soft drink consumed daily during the summer, he or she would probably find a significant relationship. However, the soft drink is not necessarily responsible for the deaths, since both variables may be related to heat and humidity.



4. There may be a *complexity of interrelationships* among many variables.

For example,

- A researcher may find a significant relationship between students' high school grades and college grades. But there probably are many other variables involved, such as IQ, hours of study, influence of parents, motivation, age, and instructors.

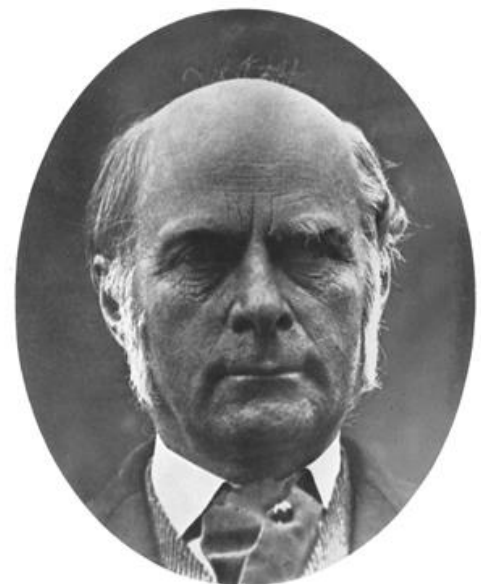


5. The relationship may be *coincidental*.

For example,

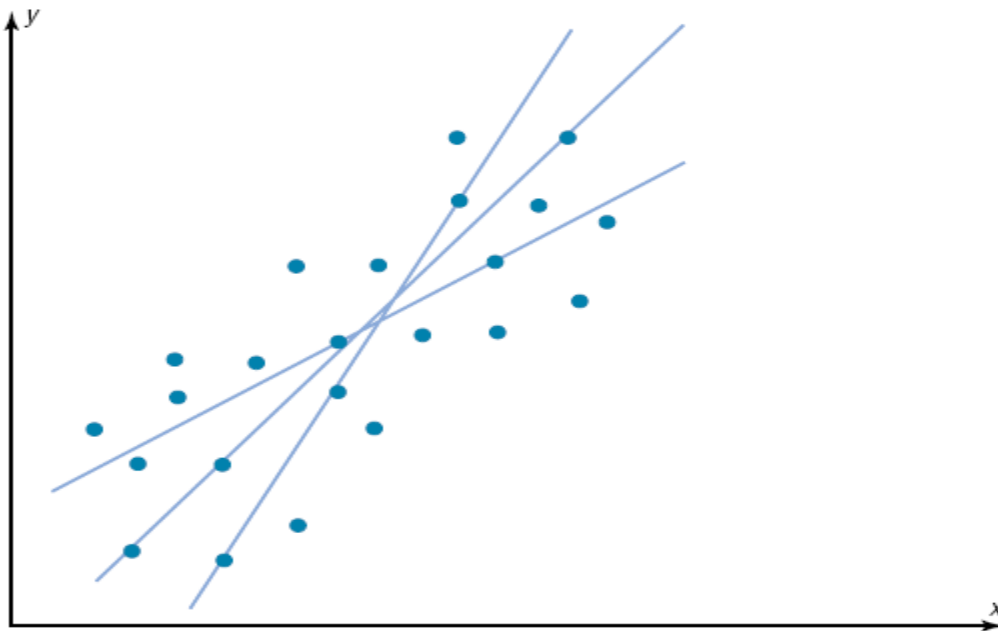
- A researcher may be able to find a significant relationship between the increase in the number of people who are exercising and the increase in the number of people who are committing crimes. But common sense dictates that any relationship between these two values must be due to coincidence.

- First introduced by Sir Francis Galton in the latter part of the 19th century
- Galton had studied the relation between heights of parents and children and noted that the heights of children of both tall and short parents appeared to “revert” or “regress” to the mean of the group.

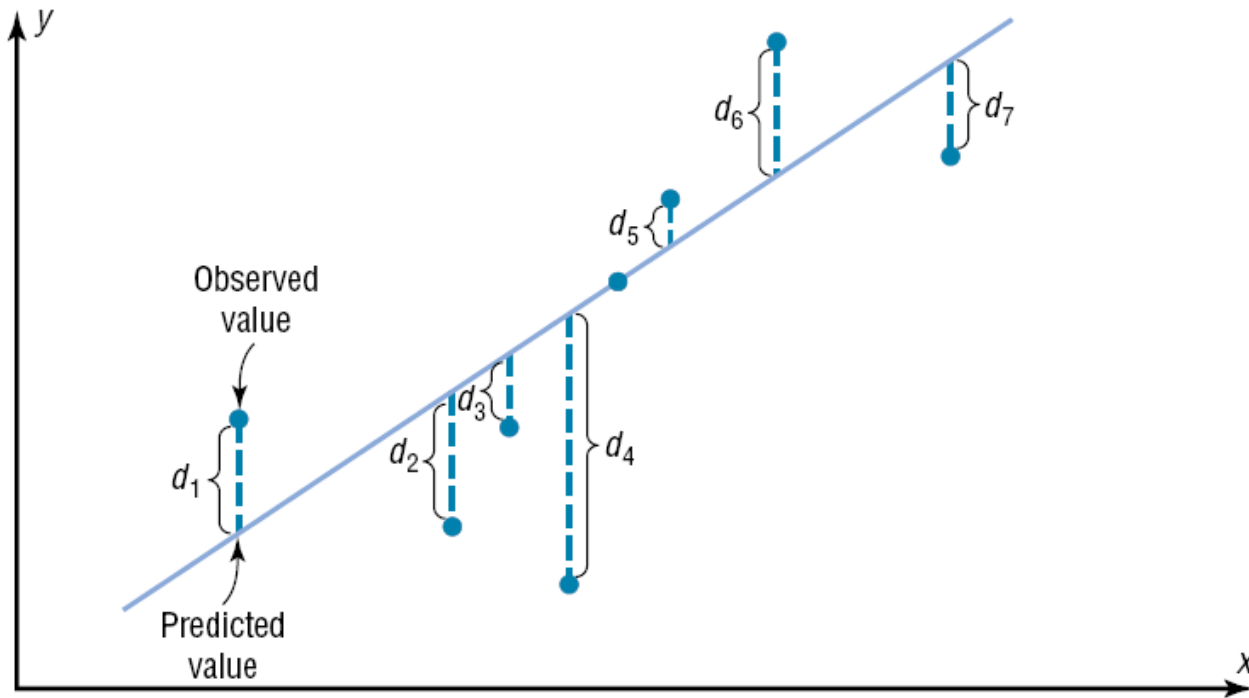


geographer, meteorologist,
tropical explorer, founder of
differential psychology, inventor
of fingerprint identification,
pioneer of statistical correlation
and regression ...

If the value of the correlation coefficient is significant, the next step is to determine the equation of the **regression line** which is the data's line of best fit.



- **Best fit** means that the sum of the squares of the vertical distance from each point to the line is at a minimum.



Online regression plot

<https://www.statskingdom.com/linear-regression-calculator.html>

$$y' = a + bx$$

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

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

where

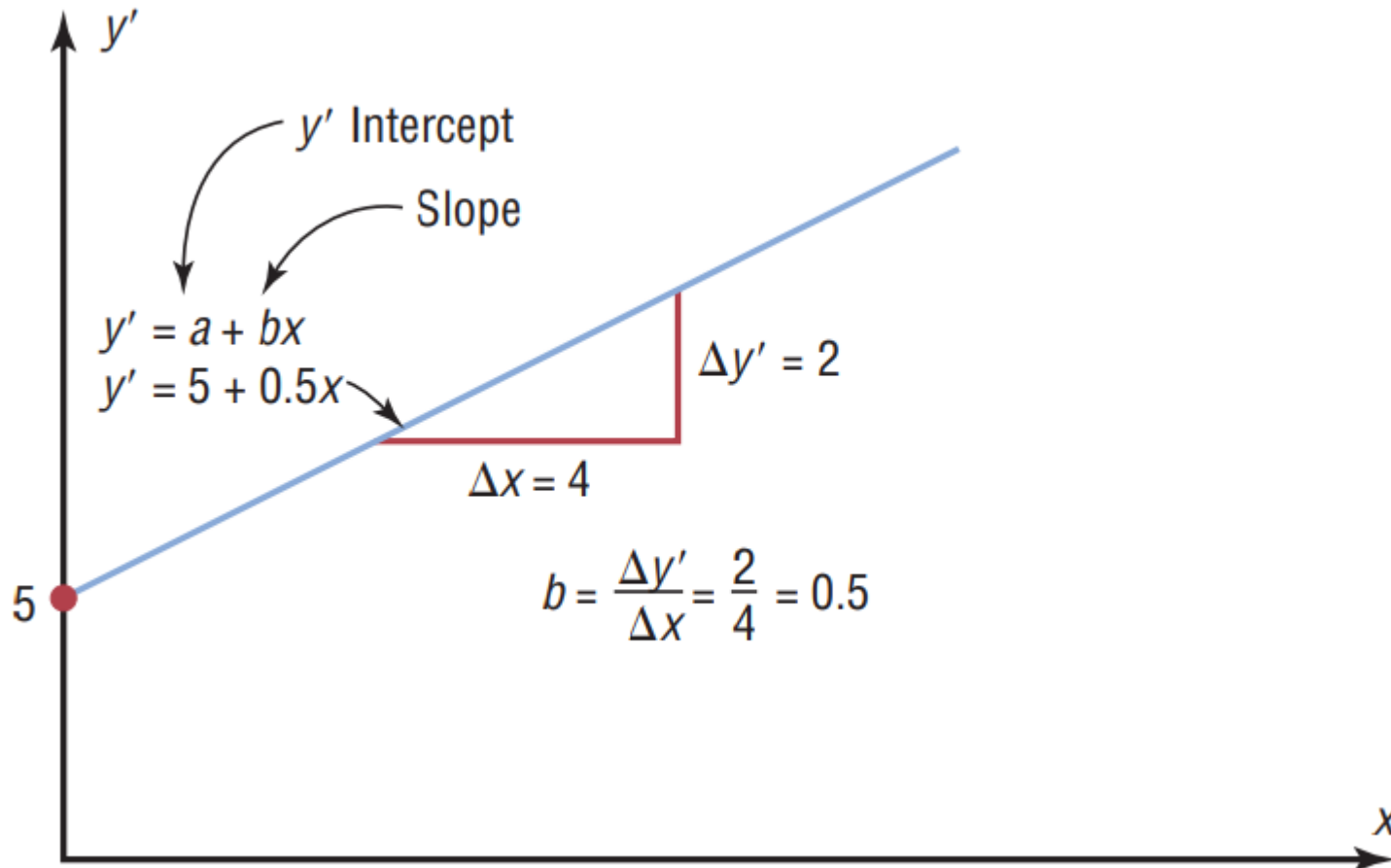
a = y' intercept

b = the slope of the line.

Regression Line



$$y' = a + bx$$



Find the equation of the regression line for the data in the car rental example, and graph the line on the scatter plot.

$$\Sigma x = 153.8, \Sigma y = 18.7, \Sigma xy = 682.77, \Sigma x^2 = 5859.26, \\ \Sigma y^2 = 80.67, n = 6$$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} \\ = \frac{(18.7)(5859.26) - (153.8)(682.77)}{6(5859.26) - (153.8)^2} = 0.396$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{6(682.77) - (153.8)(18.7)}{6(5859.26) - (153.8)^2} = 0.106$$

$$y' = a + bx \rightarrow y' = 0.396 + 0.106x$$

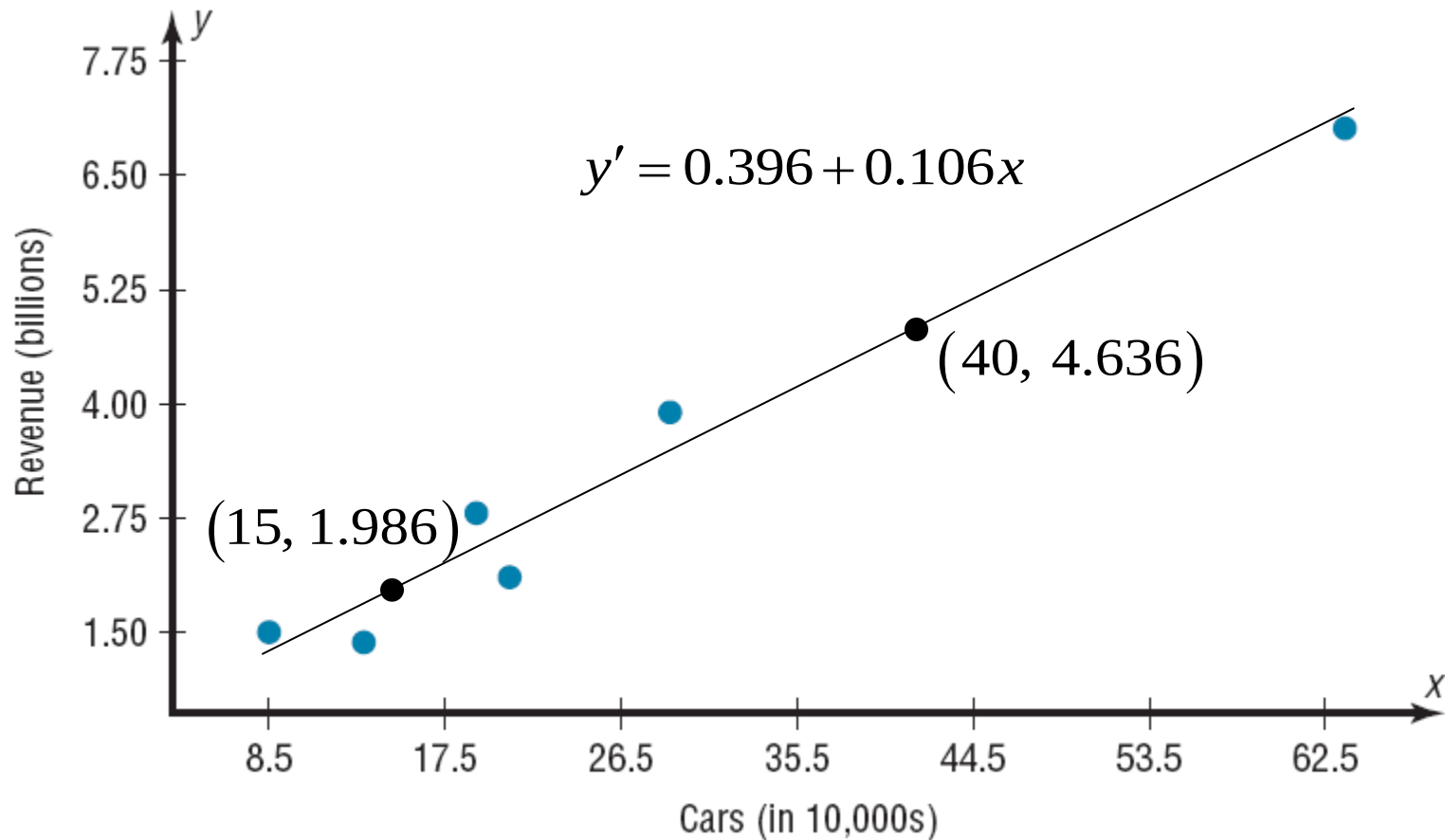
Find two points to sketch the graph of the regression line.

Use any x values between 10 and 60. For example, let x equal 15 and 40. Substitute in the equation and find the corresponding y value.

$$\begin{array}{ll} y' = 0.396 + 0.106x & y' = 0.396 + 0.106x \\ = 0.396 + 0.106(15) & = 0.396 + 0.106(40) \\ = 1.986 & = 4.636 \end{array}$$

Plot (15,1.986) and (40,4.636), and sketch the resulting line.

Example: Car Rental Companies





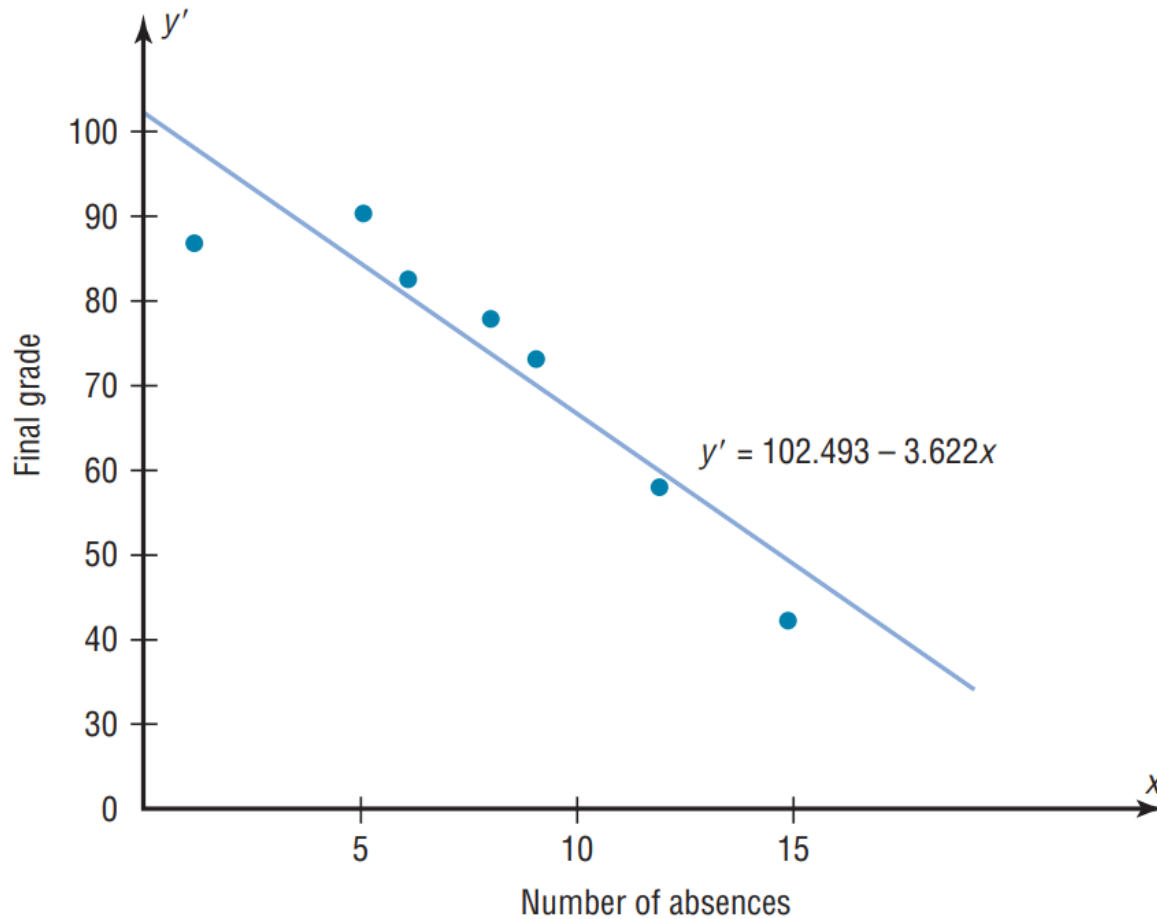
Use the equation of the regression line to predict the income of a car rental agency that has 200,000 automobiles.

$x = 20$ corresponds to 200,000 automobiles.

$$\begin{aligned}y' &= 0.396 + 0.106x \\&= 0.396 + 0.106(20) \\&= 2.516\end{aligned}$$

Hence, when a rental agency has 200,000 automobiles, its revenue will be approximately \$2.516 billion.

Example: Absences/Final Grades



When a student has 13 absences, his/her final grade will be approximately 55.407



The magnitude of the change in one variable when the other variable changes exactly 1 unit is called a **marginal change**. The value of slope b of the regression line equation represents the marginal change.

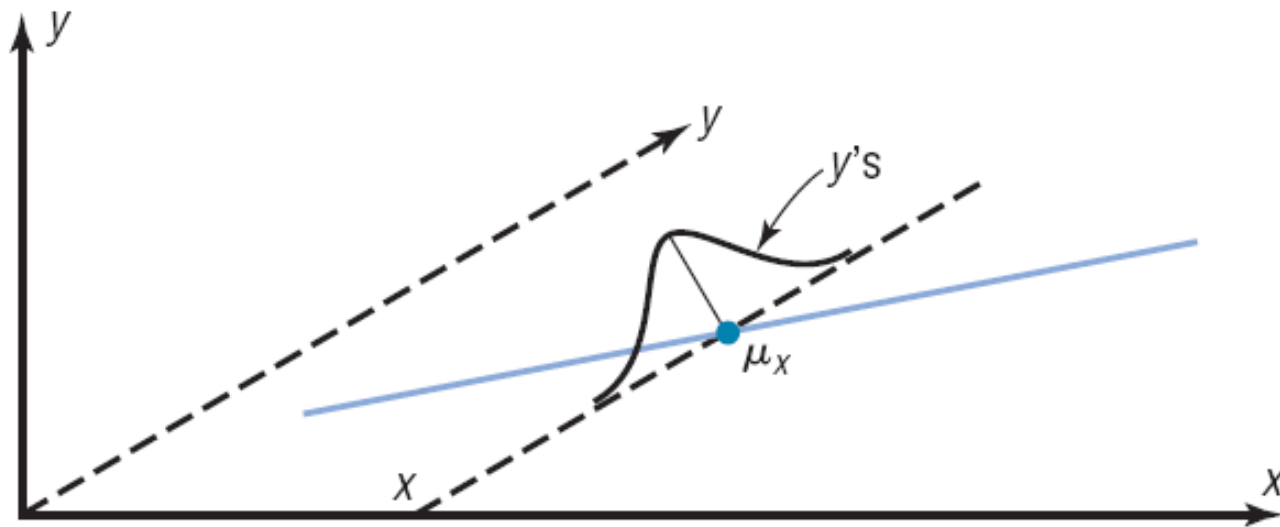
For valid predictions, the value of the correlation coefficient must be significant.

When r is not significantly different from 0, the best predictor of y is the mean of the data values of y .

Assumptions for Valid Predictions

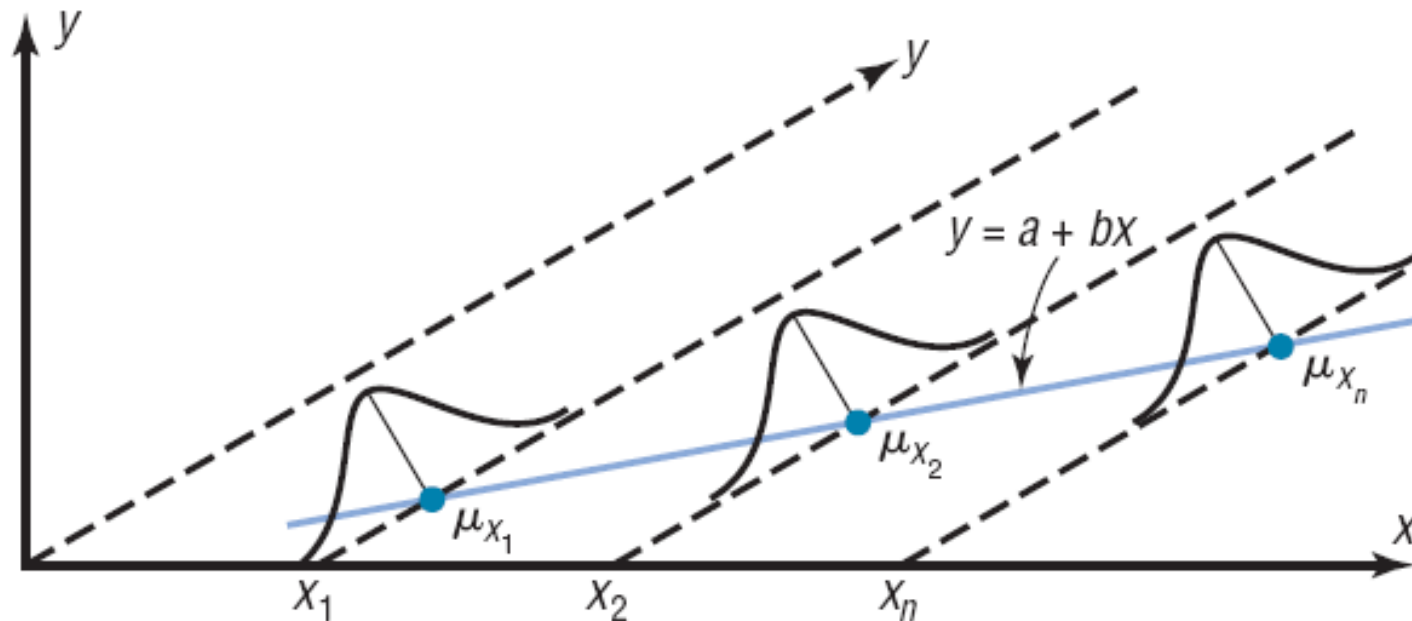


1. For any specific value of the independent variable x , the value of the dependent variable y must be normally distributed about the regression line.



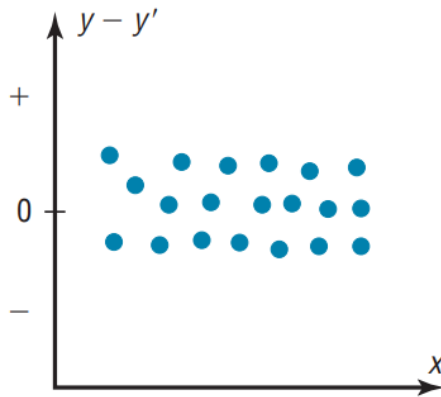
(a) Dependent variable y normally distributed

2. The standard deviation of each of the dependent variables must be the same for each value of the independent variable.

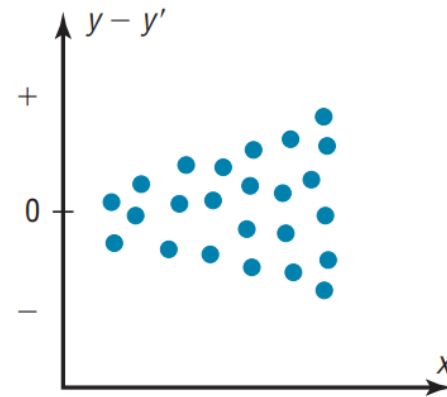


(b) $\sigma_1 = \sigma_2 = \dots = \sigma_n$

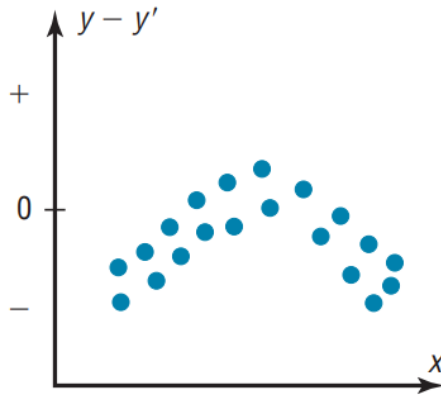
Residual Plot



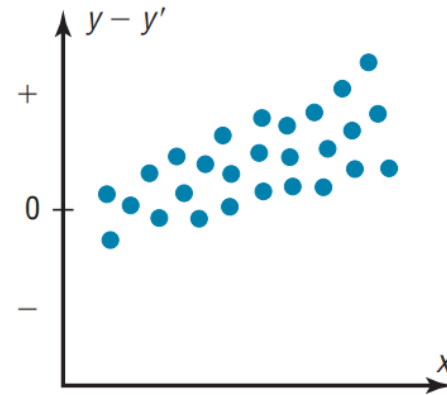
(a)



(b)



(c)



(d)

The regression line can be questionable for making predictions if assumptions do not hold.



Extrapolation, or making predictions beyond the bounds of the data, must be interpreted cautiously.

Remember that when predictions are made, they are based on present conditions or on the premise that present trends will continue. This assumption may or may not prove true in the future.



The **total variation** $\sum (y - \bar{y})^2$ is the sum of the squares of the vertical distances each point is from the mean.

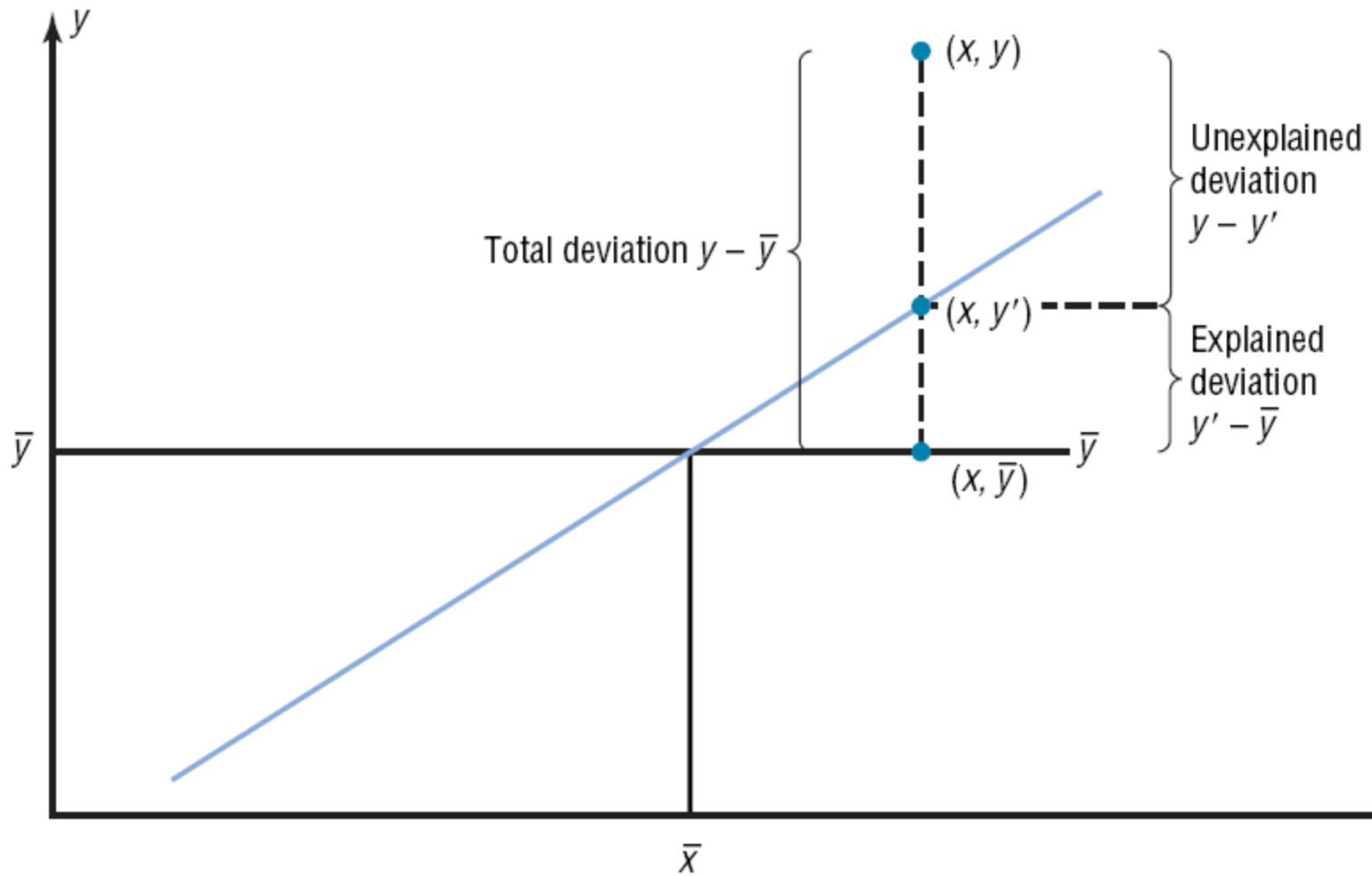
The total variation can be divided into two parts: that which is attributed to the relationship of x and y , and that which is due to chance.

The variation obtained from the relationship

$\sum (y' - \bar{y})^2$ (i.e., from the predicted y' values) is and is called the **explained variation**.

Variation due to chance, found by $\sum (y' - y)^2$, is called the **unexplained variation**. This variation cannot be attributed to the relationships.

Variation





The **coefficient of determination** is the ratio of the explained variation to the total variation.

The symbol for the coefficient of determination is r^2 .

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

Another way to arrive at the value for r^2 is to square the correlation coefficient.



The **standard error of the estimate**, denoted by s_{est} is the standard deviation of the observed y values about the predicted y' values. The formula for the standard error of estimate is:

$$s_{est} = \sqrt{\frac{\sum (y - y')^2}{n - 2}}$$

Example: Copy Machine Costs



A researcher collects the following data and determines that there is a significant relationship between the age of a copy machine and its monthly maintenance cost. The regression equation is $y' = 55.57 + 8.13x$. Find the standard error of the estimate.

Machine	Age x (years)	Monthly cost y
A	1	\$ 62
B	2	78
C	3	70
D	4	90
E	4	93
F	6	103

Example: Copy Machine Costs



Machine	Age x (years)	Monthly cost, y	y'	$y - y'$	$(y - y')^2$
A	1	62	63.70	-1.70	2.89
B	2	78	71.83	6.17	38.0689
C	3	70	79.96	-9.96	99.2016
D	4	90	88.09	1.91	3.6481
E	4	93	88.09	4.91	24.1081
F	6	103	104.35	-1.35	1.8225
					169.7392

$$y' = 55.57 + 8.13x$$

$$y' = 55.57 + 8.13(1) = 63.70$$

$$y' = 55.57 + 8.13(2) = 71.83$$

$$y' = 55.57 + 8.13(3) = 79.96$$

$$y' = 55.57 + 8.13(4) = 88.09$$

$$y' = 55.57 + 8.13(6) = 104.35$$

$$s_{est} = \sqrt{\frac{\sum (y - y')^2}{n - 2}}$$

$$s_{est} = \sqrt{\frac{169.7392}{4}} = 6.51$$

Another Formula to compute s_{est}

$$s_{est} = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}}$$

Example: Copy Machine Costs



x	y	xy	y^2
1	62	62	3,844
2	78	156	6,084
3	70	210	4,900
4	90	360	8,100
4	93	372	8,649
6	<u>103</u>	<u>618</u>	<u>10,609</u>
	$\Sigma y = 496$	$\Sigma xy = 1778$	$\Sigma y^2 = 42,186$

$$s_{est} = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}}$$

$$s_{est} = \sqrt{\frac{42,186 - 55.57(496) - 8.13(1778)}{4}} = 6.48$$

$$y' - t_{\alpha/2} s_{est} \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n \sum x^2 - (\sum x)^2}} < y$$
$$< y' + t_{\alpha/2} s_{est} \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n \sum x^2 - (\sum x)^2}}$$

with d.f. = $n - 2$

find the 95% prediction interval for the monthly maintenance cost of a machine that is 3 years old.

$$\begin{aligned} 79.96 - (2.776)(6.48) \sqrt{1 + \frac{1}{6} + \frac{6(3 - 3.3)^2}{6(82) - (20)^2}} &< y \\ &< 79.96 + (2.776)(6.48) \sqrt{1 + \frac{1}{6} + \frac{6(3 - 3.3)^2}{6(82) - (20)^2}} \\ 79.96 - 19.43 &< y < 79.96 + 19.43 \\ 60.53 &< y < 99.39 \end{aligned}$$

Hence, you can be 95% confident that the interval $60.53 < y < 99.39$ contains the actual value of y .



The previous sections explained the concepts of simple linear regression and correlation. In simple linear regression, the regression equation contains one independent variable x and one dependent variable y and is written as

$$y' = a + bx$$

In multiple regression, there are several independent variables and one dependent variable, and the equation is

$$y' = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

where , $x_1, x_2, \dots, x_k$ are the independent variables.