Statistical Inference
Regression and Correlation

Linear regression and correlation are two commonly used methods

- for examining the relationship between quantitative variables and
- for making predictions regression equation,
- the equation of the line that best fits a set of data points.

Coefficient of determination,

- a descriptive measure of the utility of the regression equation for making predictions linear correlation coefficient,
- it provides a descriptive measure of the strength of the linear relationship between two quantitative variables.

Scatterplot (or **scatter diagram**)
A graph of data from two quantitative variables of a population

To construct a scatterplot
We use a horizontal axis for the observations of one variable and a vertical axis for the observations of the other. Each pair of observations is then plotted as a point.

Error, e, is the signed vertical distance from the line to a data point.
$e = y - \hat{y}$

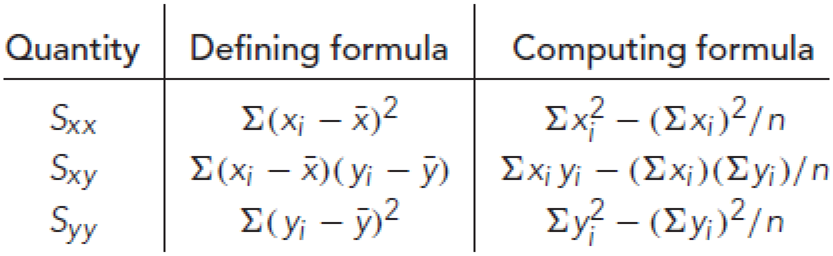
Least-Squares Criterion
Because we could draw many different lines through the cluster of data points, we need a method to choose the “best” line. The method, called the least-squares criterion, is based on an analysis of the errors made in using a line to fit the data points.

The **least-squares criterion** is that the line that best fits a set of data points is the one having the smallest possible sum of squared errors.

Regression line: The line that best fits a set of data points according to the least-squares criterion.
Regression equation: The equation of the regression line.

Response variable: The variable to be measured or observed.
Predictor variable: A variable used to predict or explain the values of the response variable.

Notation Used in Regression and Correlation
For a set of n data points, the defining and computing formulas for **Sxx, Sxy,** and **Syy** are as follows.



Quantity	Defining formula	Computing formula
S_{xx}	$\Sigma(x_i - \bar{x})^2$	$\Sigma x_i^2 - (\Sigma x_i)^2/n$
S_{xy}	$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n$
S_{yy}	$\Sigma(y_i - \bar{y})^2$	$\Sigma y_i^2 - (\Sigma y_i)^2/n$

Regression Equation

The regression equation for a set of n data points is ŷ = b0 + b1x, where

$$b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \frac{1}{n}(\Sigma y_i - b_1 \Sigma x_i) = \bar{y} - b_1\bar{x}.$$

Example:

Age and Price of Orions

a. Determine the regression equation for the data.

b. Graph the regression equation and the data points.

c. Describe the apparent relationship between age and price of Orions.

d. Interpret the slope of the regression line in terms of prices for Orions.

e. Use the regression equation to predict the price of a 3-year-old Orion and a 4-year-old Orion.

Age (yr)	**Price ($100)**		
x	**y**	**xy**	**x^2**
5	85	425	25
4	103	412	16
6	70	420	36
5	82	410	25
5	89	445	25
5	98	490	25
6	66	396	36
6	95	570	36
2	169	338	4
7	70	490	49
7	48	336	49
58	975	4732	326

Solution

a. We first need to compute b1 and b0. We did so by constructing a table of values for x (age), y (price), xy, x2, and their sums in Table.

The slope of the regression line therefore is

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n} = \frac{4732 - (58)(975)/11}{326 - (58)^2/11} = -20.26.$$

The y-intercept is

$$b_0 = \frac{1}{n}(\Sigma y_i - b_1 \Sigma x_i) = \frac{1}{11}\left[975 - (-20.26) \cdot 58\right] = 195.47.$$

So the regression equation is

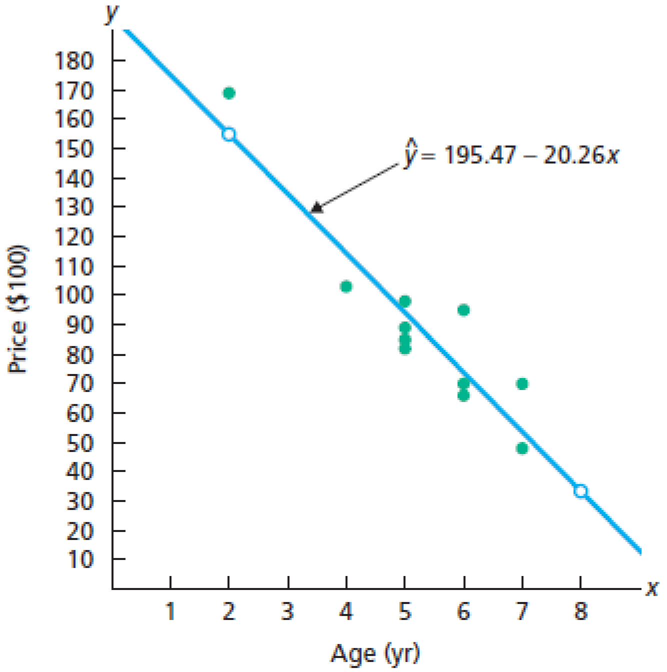
$$\hat{y} = 195.47 - 20.26x.$$

Note: The usual warnings about rounding apply. When computing the slope, b1, of the regression line, do not round until the computation is finished. When computing the y-intercept, b0, do not use the rounded value of b1; instead, keep full calculator accuracy.

b. To graph the regression equation, we need to substitute two different x-values in the regression equation to obtain two distinct points. Let's use the x-values 2 and 8. The corresponding y-values are

$$\hat{y} = 195.47 - 20.26 \cdot 2 = 154.95 \quad \text{and} \quad \hat{y} = 195.47 - 20.26 \cdot 8 = 33.39.$$

Therefore, the regression line goes through the two points (2, 154.95) and (8, 33.39). In Fig. below we plotted these two points with open dots. Drawing a line through the two open dots yields the regression line, the graph of the regression equation. The Figure also shows the data points from the first two columns of Table.



c. Because the slope of the regression line is negative, price tends to decrease as age increases, which is no particular surprise.

d. Because x represents age in years and y represents price in hundreds of dollars, the slope of −20.26 indicates that Orions depreciate an estimated $2026 per year, at least in the 2- to 7-year-old range.

e. For a 3-year-old Orion, x = 3, and the regression equation yields the predicted price of

$$\hat{y} = 195.47 - 20.26 \cdot 3 = 134.69.$$

Similarly, the predicted price for a 4-year-old Orion is

$$\hat{y} = 195.47 - 20.26 \cdot 4 = 114.43.$$

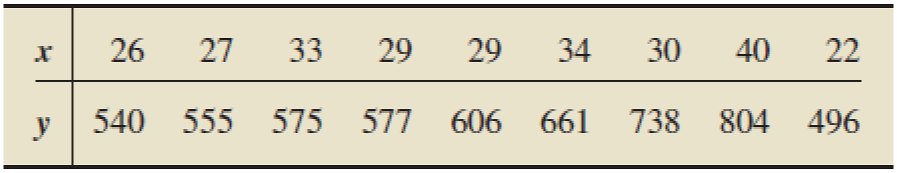
Interpretation The estimated price of a 3-year-old Orion is $13,469, and the estimated price of a 4-year-old Orion is $11,443.

Exercise:

For each of the given questions

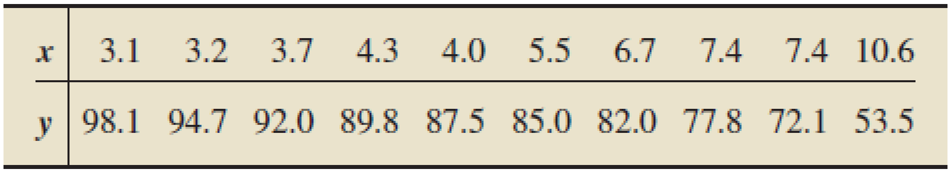
a. find the regression equation for the data points.
b. graph the regression equation and the data points.
c. describe the apparent relationship between the two variables under consideration.
d. interpret the slope of the regression line.
e. identify the predictor and response variables.
g. predict the values of the response variable for the specified values of the predictor variable, and interpret your results.

1. **Custom Homes.** Hanna Properties specializes in custom home resales in the Equestrian Estates, an exclusive subdivision in Phoenix, Arizona. A random sample of nine custom homes currently listed for sale provided the following information on size and price. Here, x denotes size, in hundreds of square feet, rounded to the nearest hundred, and y denotes price, in thousands of dollars, rounded to the nearest thousand. For part (g), predict the price of a 2600-sq. ft. home in the Equestrian Estates.



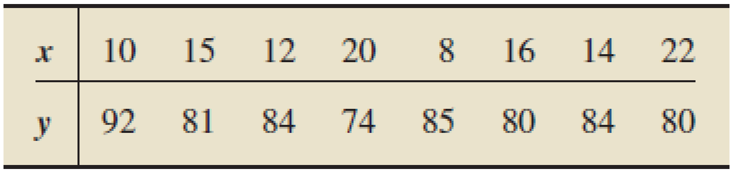
x	26	27	33	29	29	34	30	40	22
y	540	555	575	577	606	661	738	804	496

2. **Tax Efficiency.** Tax efficiency is a measure, ranging from 0 to 100, of how much tax due to capital gains stock or mutual funds investors pay on their investments each year; the higher the tax efficiency, the lower is the tax. In the article “At the Mercy of the Manager” (Financial Planning, Vol. 30(5), pp. 54–56), C. Israelsen examined the relationship between investments in mutual fund portfolios and their associated tax efficiencies. The following table shows percentage of investments in energy securities (x) and tax efficiency (y) for 10 mutual fund portfolios. For part (g), predict the tax efficiency of a mutual fund portfolio with 5.0% of its investments in energy securities and one with 7.4% of its investments in energy securities.



x	3.1	3.2	3.7	4.3	4.0	5.5	6.7	7.4	7.4	10.6
y	98.1	94.7	92.0	89.8	87.5	85.0	82.0	77.8	72.1	53.5

3. **Study Time and Score.** An instructor at Arizona State University asked a random sample of eight students to record their study times in a beginning calculus course. She then made a table for total hours studied (x) over 2 weeks and test score (y) at the end of the 2 weeks. Here are the results. For part (g), predict the score of a student who studies for 15 hours.



x	10	15	12	20	8	16	14	22
y	92	81	84	74	85	80	84	80

The Coefficient of Determination

Determine the percentage of variation in the observed values of the response variable that is explained by the regression (or predictor variable)

To find this percentage, we need to define two measures of variation:
(1) the total variation in the observed values of the response variable and
(2) the amount of variation in the observed values of the response variable that is explained by the regression.

Sums of Squares in Regression

Total sum of squares, SST:
The total variation in the observed values of the response variable:

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

Regression sum of squares, SSR:
The variation in the observed values of the response variable explained by the regression:

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

Error sum of squares, SSE:
The variation in the observed values of the response variable not explained by the regression:

$$SSE = \Sigma(y_i - \hat{y}_i)^2$$

Coefficient of Determination

The **coefficient of determination, r^2,** is the proportion of variation in the observed values of the response variable explained by the regression. Thus,

$$r^2 = \frac{SSR}{SST}.$$

The coefficient of determination is a descriptive measure of the utility of the regression equation for making predictions.

Note: The coefficient of determination, r^2, always lies between 0 and 1.
A value of r^2 near 0 suggests that the regression equation is not very useful for making predictions, whereas a value of r^2 near 1 suggests that the regression equation is quite useful for making predictions.

Regression Identity

The total sum of squares equals the regression sum of squares plus the error sum of squares:
SST = SSR + SSE.

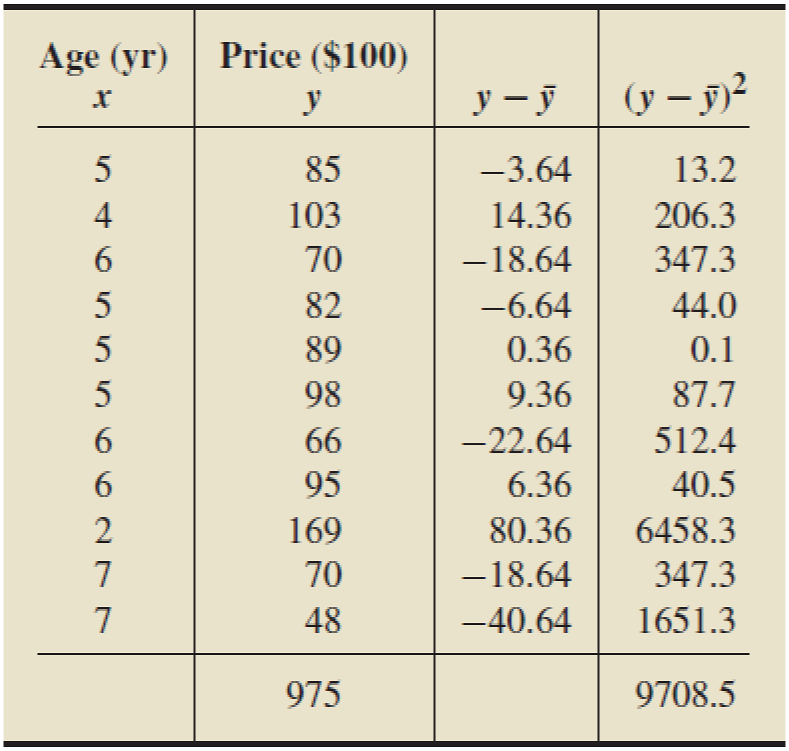
$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

EXAMPLE

The Coefficient of Determination Consider Age and Price of Orions data

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{975}{11} = 88.64.$$

Table for computing SST for the Orion price data

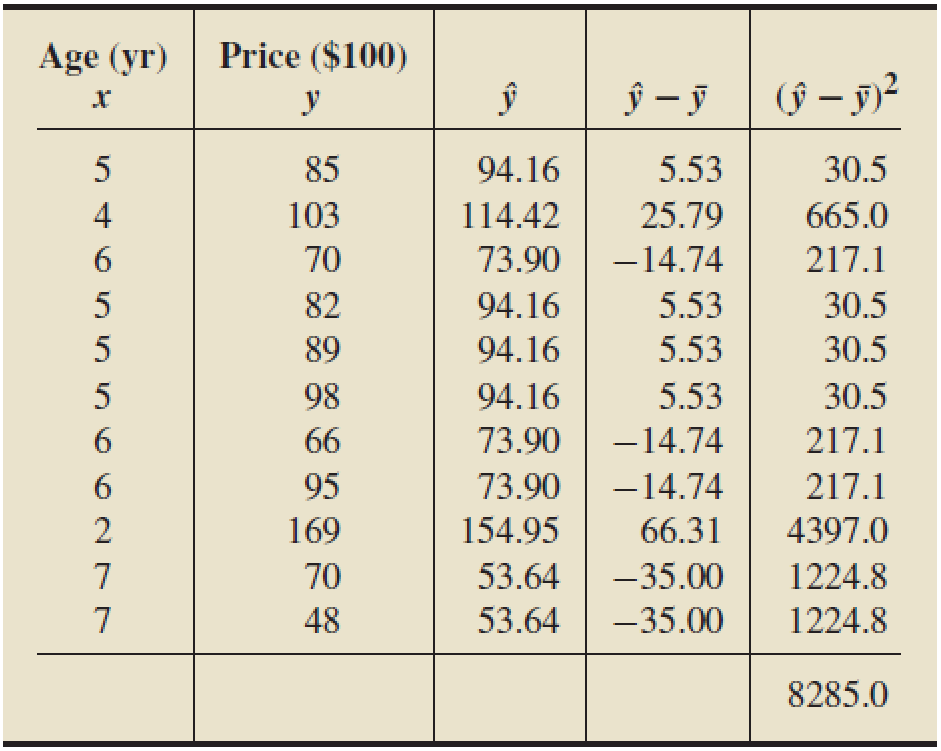


Age (yr) x	Price ($100) y	$y - \bar{y}$	$(y - \bar{y})^2$
5	85	−3.64	13.2
4	103	14.36	206.3
6	70	−18.64	347.3
5	82	−6.64	44.0
5	89	0.36	0.1
5	98	9.36	87.7
6	66	−22.64	512.4
6	95	6.36	40.5
2	169	80.36	6458.3
7	70	−18.64	347.3
7	48	−40.64	1651.3
	975		9708.5

$$SST = \Sigma(y_i - \bar{y})^2 = 9708.5.$$

Table for computing SSR for the Orion data

$$\hat{y} = 195.47 - 20.26x$$



Age (yr) x	Price ($100) y	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
5	85	94.16	5.53	30.5
4	103	114.42	25.79	665.0
6	70	73.90	−14.74	217.1
5	82	94.16	5.53	30.5
5	89	94.16	5.53	30.5
5	98	94.16	5.53	30.5
6	66	73.90	−14.74	217.1
6	95	73.90	−14.74	217.1
2	169	154.95	66.31	4397.0
7	70	53.64	−35.00	1224.8
7	48	53.64	−35.00	1224.8
				8285.0

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 = 8285.0.$$

From SST and SSR, we compute the coefficient of determination, the percentage of variation in the observed prices explained by the regression (i.e., by the linear relationship between age and price for the sampled Orions):

$$r^2 = \frac{SSR}{SST} = \frac{8285.0}{9708.5} = 0.853 \quad (85.3\%).$$

Interpretation Evidently, age is quite useful for predicting price because 85.3% of the variation in the observed prices is explained by the regression of price on age.

The Regression Identity
For the Orion data, SST = 9708.5, SSR = 8285.0, and SSE = 1423.5.
Because 9708.5 = 8285.0 + 1423.5, we see that SST = SSR + SSE.
This equation is always true and is called the **regression identity.**

Exercise

For the given questions

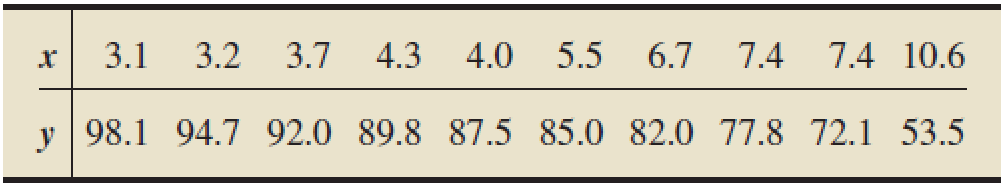
a. Compute SST, SSR, and SSE.

b. Compute the coefficient of determination, r^2.

c. Determine the percentage of variation in the observed values of the response variable explained by the regression and interpret your answer.

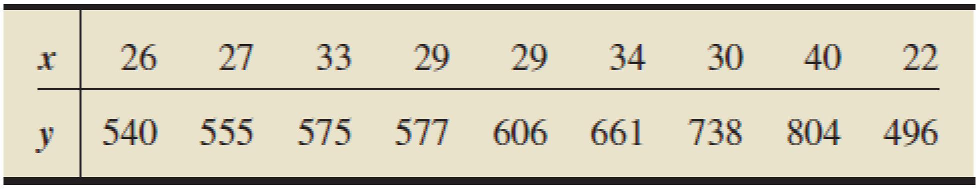
d. State how useful the regression equation appears to be for making predictions.

1. **Tax Efficiency.** Following are the data on percentage of investments in energy securities and tax efficiency



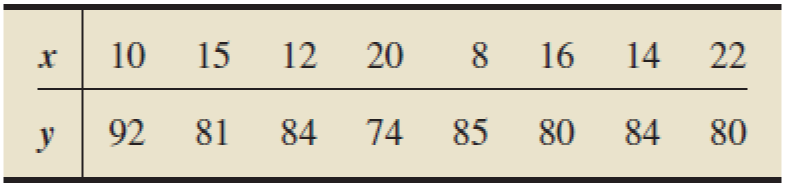
x	3.1	3.2	3.7	4.3	4.0	5.5	6.7	7.4	7.4	10.6
y	98.1	94.7	92.0	89.8	87.5	85.0	82.0	77.8	72.1	53.5

2. **Custom Homes.** Following are the size and price data for custom homes



x	26	27	33	29	29	34	30	40	22
y	540	555	575	577	606	661	738	804	496

3. **Study Time and Score.** Following are the data on study time and score for calculus students



x	10	15	12	20	8	16	14	22
y	92	81	84	74	85	80	84	80

Linear Correlation

The linear correlation coefficient is a descriptive measure of the strength and direction of the linear (straight-line) relationship between two variables.

Computing Formula for a Linear Correlation Coefficient

The computing formula for a linear correlation coefficient is

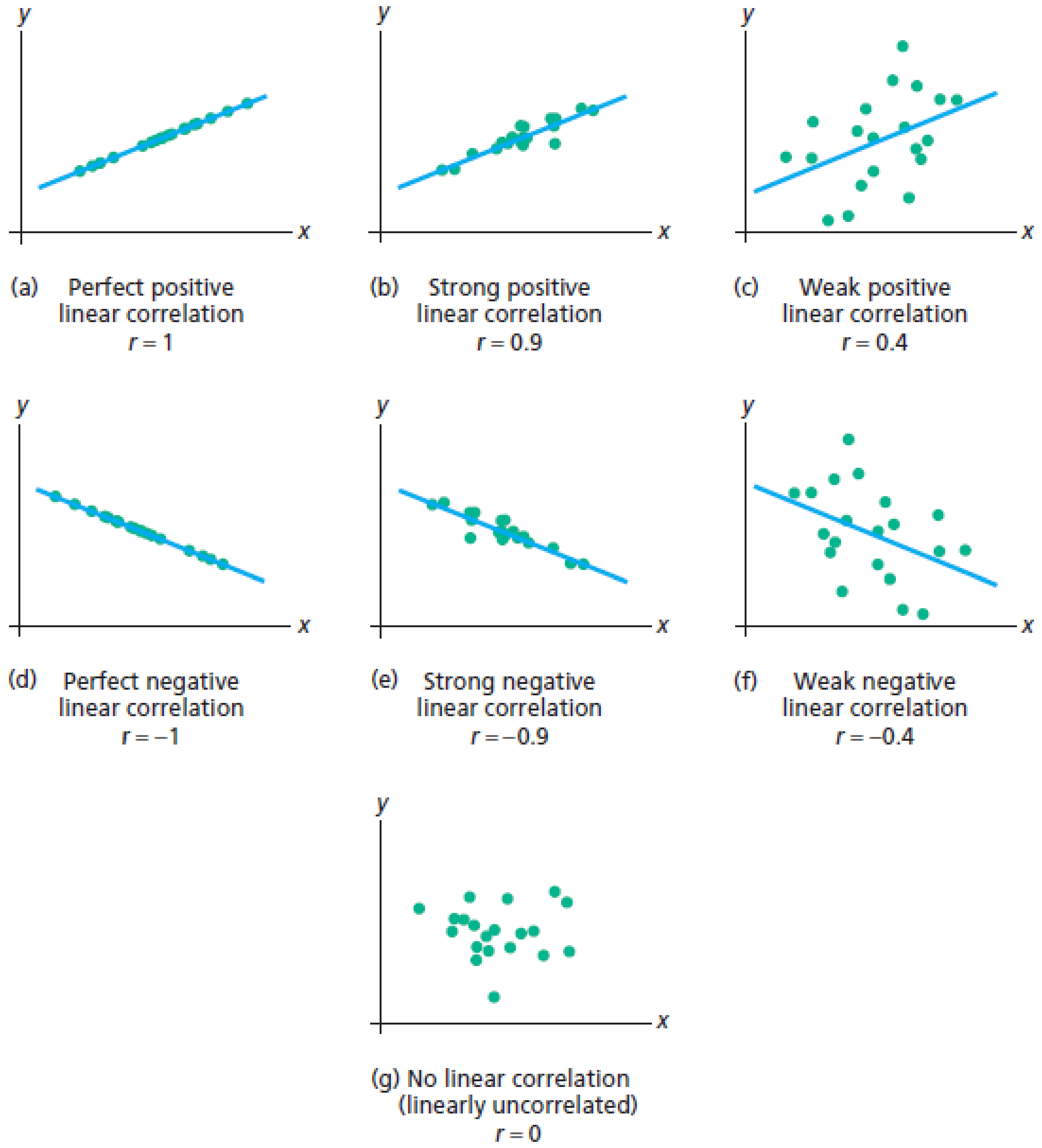
$$r = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{[\Sigma x_i^2 - (\Sigma x_i)^2/n][\Sigma y_i^2 - (\Sigma y_i)^2/n]}}$$

Or

$$r = S_{xy}/\sqrt{S_{xx}S_{yy}}$$

r is independent of the choice of units and always lies between −1 and 1.

Various degrees of linear correlation



(a) Perfect positive linear correlation $r = 1$

(b) Strong positive linear correlation $r = 0.9$

(c) Weak positive linear correlation $r = 0.4$

(d) Perfect negative linear correlation $r = -1$

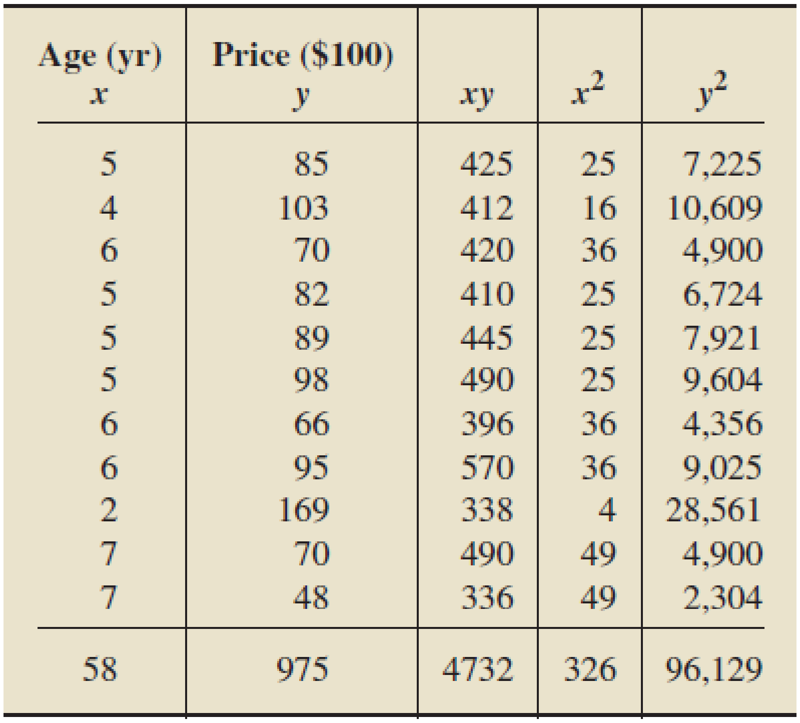
(e) Strong negative linear correlation $r = -0.9$

(f) Weak negative linear correlation $r = -0.4$

(g) No linear correlation (linearly uncorrelated) $r = 0$

EXAMPLE: The Linear Correlation Coefficient

Age and Price of Orions The age and price data for a sample of 11 Orions are repeated in the first two columns of Table



Age (yr) x	Price ($100) y	xy	x^2	y^2
5	85	425	25	7,225
4	103	412	16	10,609
6	70	420	36	4,900
5	82	410	25	6,724
5	89	445	25	7,921
5	98	490	25	9,604
6	66	396	36	4,356
6	95	570	36	9,025
2	169	338	4	28,561
7	70	490	49	4,900
7	48	336	49	2,304
58	975	4732	326	96,129

a. Compute the linear correlation coefficient, r , of the data.

b. Interpret the value of r obtained in part (a) in terms of the linear relationship between the variables age and price of Orions.

c. Discuss the graphical implications of the value of r .

Solution:

$$r = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{[\Sigma x_i^2 - (\Sigma x_i)^2/n][\Sigma y_i^2 - (\Sigma y_i)^2/n]}}$$

$$= \frac{4732 - (58)(975)/11}{\sqrt{[326 - (58)^2/11][96{,}129 - (975)^2/11]}} = -0.924.$$

b. Interpretation The linear correlation coefficient, r = −0.924, suggests a strong negative linear correlation between age and price of Orions. In particular, it indicates that as age increases, there is a strong tendency for price to decrease, which is not surprising. It also implies that the regression equation,
$\hat{y}$ = 195.47 − 20.26x, is extremely useful for making predictions.

c. Because the correlation coefficient, r = −0.924, is quite close to −1, the data points should be clustered closely about the regression line

Relationship between the Correlation Coefficient and the Coefficient of Determination
The coefficient of determination equals the square of the linear correlation coefficient.

Exercise
For each exercise here,
a. Obtain the linear correlation coefficient.
b. Interpret the value of r in terms of the linear relationship between the two variables in question.

1. **Custom Homes.** Following are the size and price data for custom homes.

x	26	27	33	29	29	34	30	40	22
y	540	555	575	577	606	661	738	804	496

2. **Plant Emissions.** Following are the data on plant weight and quantity of volatile emissions from.

x	57	85	57	65	52	67	62	80	77	53	68
y	8.0	22.0	10.5	22.5	12.0	11.5	7.5	13.0	16.5	21.0	12.0

3. **Crown-Rump Length.** Following are the data on age and crown-rump length for fetuses

x	10	10	13	13	18	19	19	23	25	28
y	66	66	108	106	161	166	177	228	235	280

Inferential Methods in Regression and Correlation

Assumptions (Conditions) for Regression Inferences

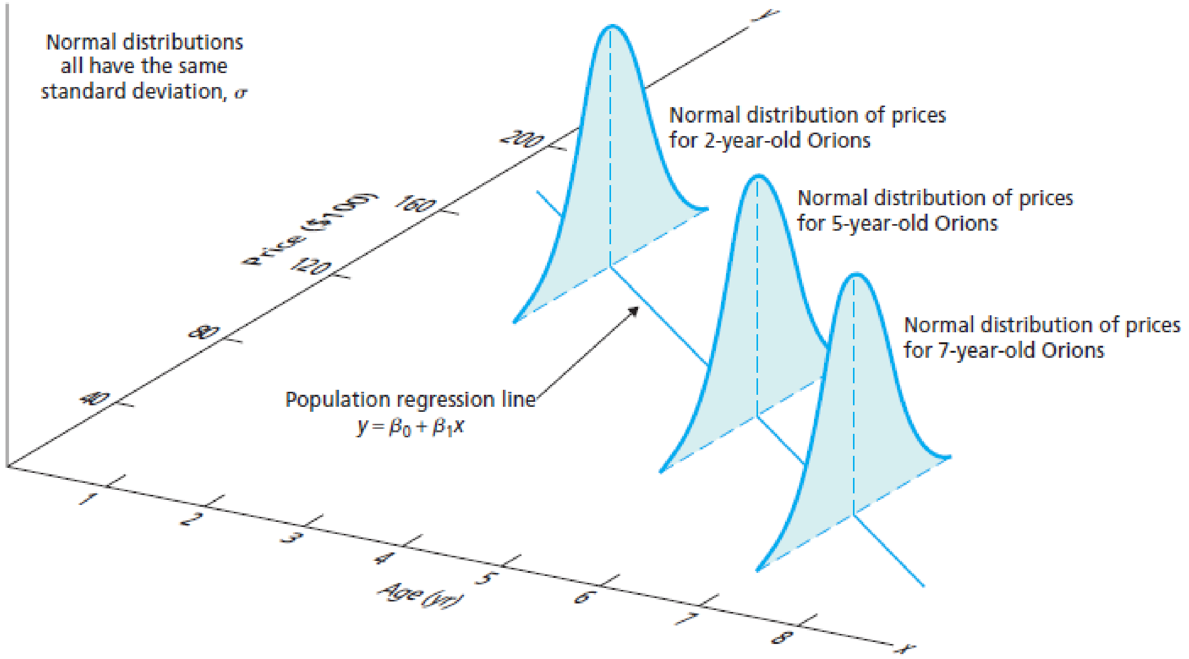
1. Population regression line: There are constants β0 and β1 such that, for each value x of the predictor variable, the conditional mean of the response variable is β0 + β1x.

2. Equal standard deviations: The conditional standard deviations of the response variable are the same for all values of the predictor variable. We denote this common standard deviation σ.

3. Normal populations: For each value of the predictor variable, the conditional distribution of the response variable is a normal distribution.

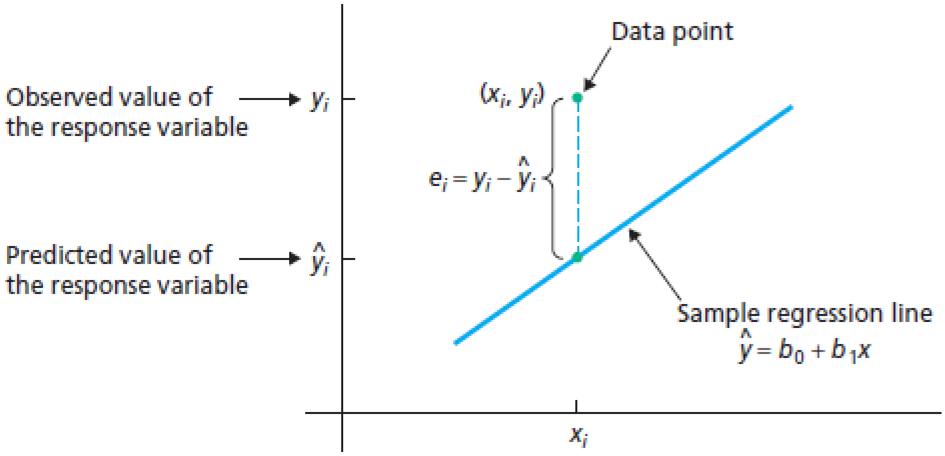
4. Independent observations: The observations of the response variable are independent of one another.

FIGURE Graphical portrayal of Assumptions 1–3 for regression inferences pertaining to age and price of Orion


Normal distributions
all have the same
standard deviation, σ
Normal distribution of prices
for 2-year-old Orions
Normal distribution of prices
for 5-year-old Orions
Normal distribution of prices
for 7-year-old Orions
Population regression line
$y = \beta_0 + \beta_1 x$
Price ($100)
40
80
120
160
200
y
1
2
3
4
5
6
7
8
Age (yr)
x

$$\text{Residual} = e_i = y_i - \hat{y}_i .$$

the residual of a single data point.



t-Distribution for Inferences for β1 Regression t-Test

Purpose To perform a hypothesis test to decide whether a predictor variable is useful for making predictions
Assumptions The four assumptions for regression inferences

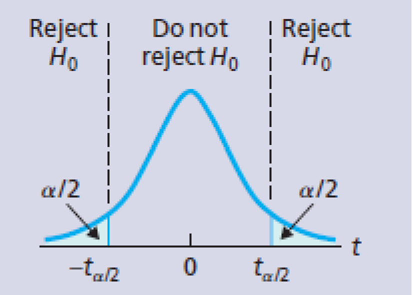
Step 1 The null and alternative hypotheses are, respectively,
H0: β1 = 0 (predictor variable is not useful for making predictions)
Ha: β1 ≠ 0 (predictor variable is useful for making predictions).
Step 2 Decide on the significance level, α.
Step 3 Compute the value of the test statistic

$$t = \frac{b_1}{s_e/\sqrt{S_{xx}}}$$

and denote that value t_0.
se = standard error of the estimate

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\Sigma e_i^2}{n-2}}.$$

Step 4 The critical values are ±tα/2 with df = n − 2. Use T Table to find the critical values.



Step 5 If the value of the test statistic falls in the rejection region, reject H0; otherwise, do not reject H0.
Step 6 Interpret the results of the hypothesis test.

EXAMPLE The Regression t-Test

Age and Price of Orions The data on age and price for a sample of 11 Orions are displayed in Table. At the 5% significance level, do the data provide sufficient evidence to conclude that age is useful as a (linear) predictor of price for Orions?

Step 1 State the null and alternative hypotheses.
Let β1 denote the slope of the population regression line that relates price to age for Orions. Then the null and alternative hypotheses are, respectively,
H_0: β1 = 0 (age is not useful for predicting price)
Ha: β1 ≠ 0 (age is useful for predicting price).
Step 2 Decide on the significance level, α.
We are to perform the hypothesis test at the 5% significance level, or α = 0.05.
Step 3 Compute the value of the test statistic

$$SST = \Sigma y_i^2 - (\Sigma y_i)^2/n = 96{,}129 - (975)^2/11 = 9708.5;$$

regression sum of squares is

$$SSR = \frac{\left[\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n\right]^2}{\Sigma x_i^2 - (\Sigma x_i)^2/n} = \frac{\left[4732 - (58)(975)/11\right]^2}{326 - (58)^2/11} = 8285.0;$$

from the two preceding results, the error sum of squares is

$$SSE = SST - SSR = 9708.5 - 8285.0 = 1423.5.$$

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1423.5}{11-2}} = 12.58.$$

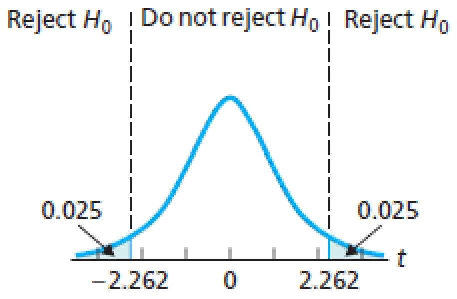
we found that $b_1 = -20.26$, $\Sigma x_i^2 = 326$, and $\Sigma x_i = 58$

Therefore, because $n = 11$, the value of the test statistic is

$$t = \frac{b_1}{s_e/\sqrt{S_{xx}}} = \frac{b_1}{s_e/\sqrt{\Sigma x_i^2 - (\Sigma x_i)^2/n}}$$

$$= \frac{-20.26}{12.58/\sqrt{326 - (58)^2/11}} = -7.235.$$

Step 4 The critical values are ±tα/2 with df = n − 2. Use T Table to find the critical values.
From Step 2, α = 0.05. For n = 11, df = n − 2 = 11 − 2 = 9. Using Table, we find that the critical values are ±tα/2 = ±t0.025 = ±2.262, as depicted in
Fig. 15.10A.



Step 5 If the value of the test statistic falls in the rejection region, reject H0; otherwise, do not reject H0.
The value of the test statistic, found in Step 3, is t = −7.235. Because this value falls in the rejection region, we reject H0. The test results are statistically significant at the 5% level.

Step 6 Interpret the results of the hypothesis test.
Interpretation At the 5% significance level, the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and hence that age is useful as a (linear) predictor of price for Orions.

Exercise

Presuming that the assumptions for regression inferences are met, decide at the specified significance level whether the data provide sufficient evidence to conclude that the predictor variable is useful for predicting the response variable.

1. **Tax Efficiency.** Following are the data on percentage of investments in energy securities and tax efficiency. Use α = 0.05.

x	3.1	3.2	3.7	4.3	4.0	5.5	6.7	7.4	7.4	10.6
y	98.1	94.7	92.0	89.8	87.5	85.0	82.0	77.8	72.1	53.5

2. **Corvette Prices.** Following are the age and price data for Corvettes. Use α = 0.10.

x	6	6	6	2	2	5	4	5	1	4
y	290	280	295	425	384	315	355	328	425	325