National University of Computer & Emerging Sciences
MT-2005 Probability and Statistics

Simple Linear Regression

- The dependence of one variable over the other variable is called “Regression”.
- The statistical method which helps us to estimate the value of dependent variable from the known value of independent variable is called regression.
- The statistical model for simple linear regression is given below. The response Y is related to the independent variable x through the equation (True regression line):

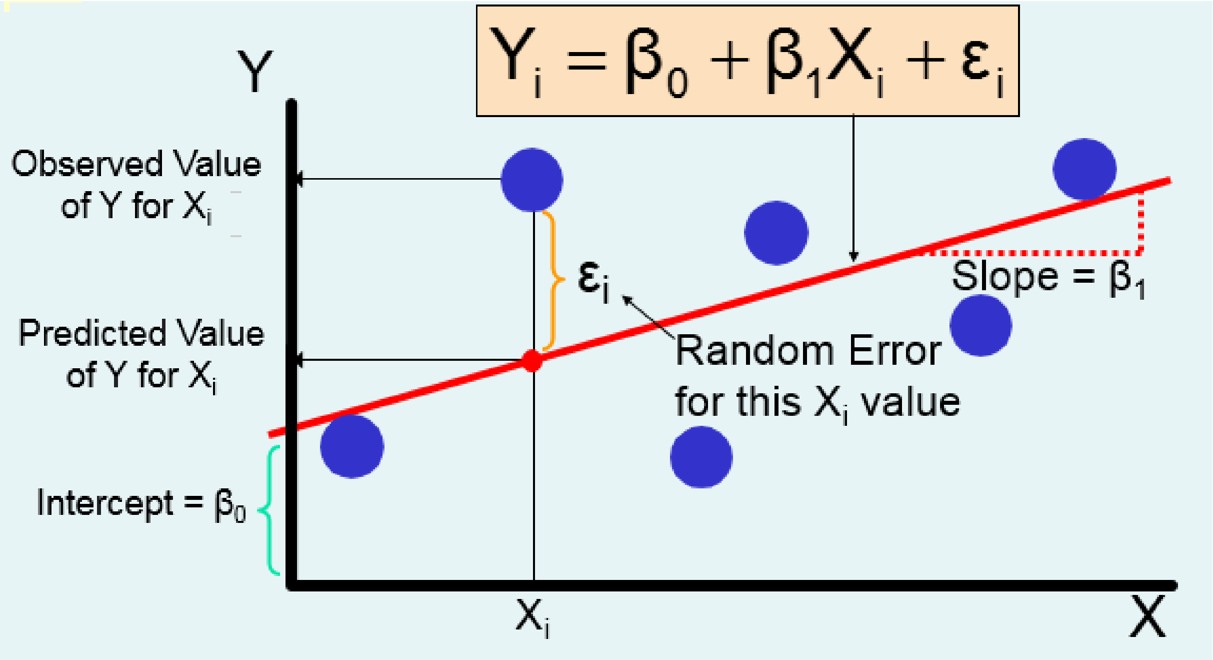
$$Y = \beta_0 + \beta_1 x + \epsilon.$$

In the above, β_0 and β_1 are unknown intercept and slope parameters, respectively, and ϵ is a random variable that is assumed to be distributed with $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. The quantity σ^2 is often called the error variance or residual variance.

- The quantity Y is a random variable since ϵ is random.
- The value x of the regressor variable is not random and, in fact, is measured with negligible error.
- The quantity ϵ, often called a random error or random disturbance, has constant variance (homogeneous variance)
- The presence of this random error, keeps the model from becoming simply a deterministic equation.

We must keep in mind that:

- In practice ß0 and ß1 are not known and must be estimated from data.
- we never observe the actual ϵ values in practice and thus we can never draw the true regression line
- We can only draw an estimated line.

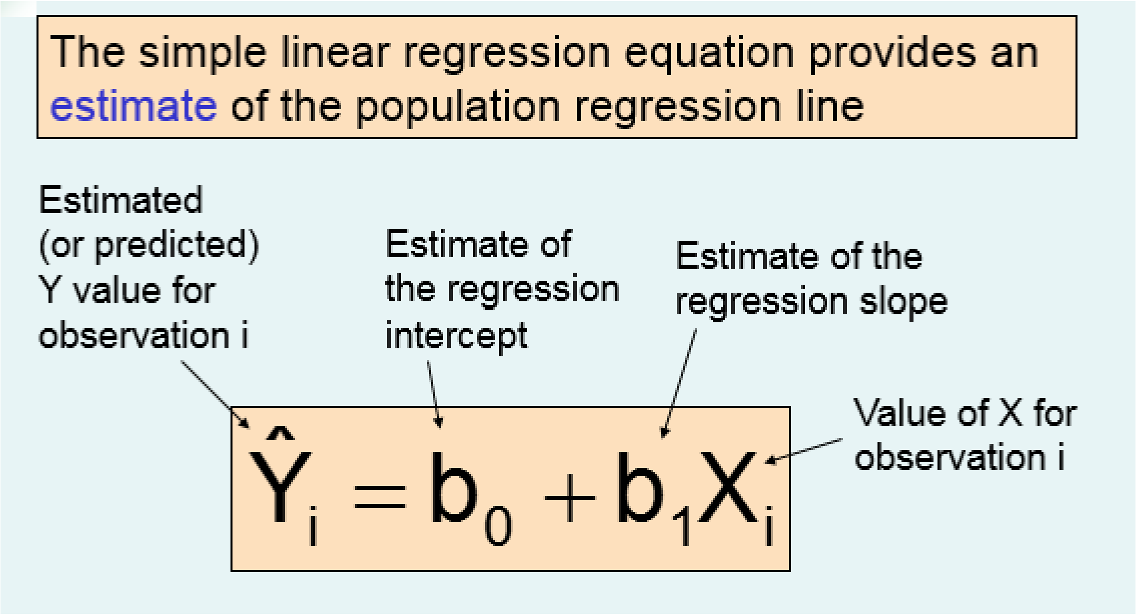

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
ε_i
Predicted Value of Y for X_i
Slope = β_1
Random Error for this X_i value
Intercept = β_0
X_i
X

Deterministic Vs. Statistical Relationship

- statistical, not deterministic dependence among variables.
- In statistical relationships among variables we essentially deal with random or stochastic variables i.e. variables that have probability distributions.
- The dependence of crop yield on temperature, rainfall, sunshine, and fertilizer, for example, is statistical in nature in the sense that the explanatory variables, although certainly important, will not enable the agronomist to predict crop yield exactly.
- In deterministic phenomena, on the other hand, we deal with relationships of the type, say, exhibited by Newton’s law of gravity, which states: Every particle in the universe attracts every other particle with a force directly proportional to the product of their masses and inversely proportional to the square of the distance between them.

$$F = k(m_1 m_2 / r^2),$$

Simple Linear Regression Equation as Estimator Of Population Regression Line:



The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i
Estimate of the regression intercept
Estimate of the regression slope
Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

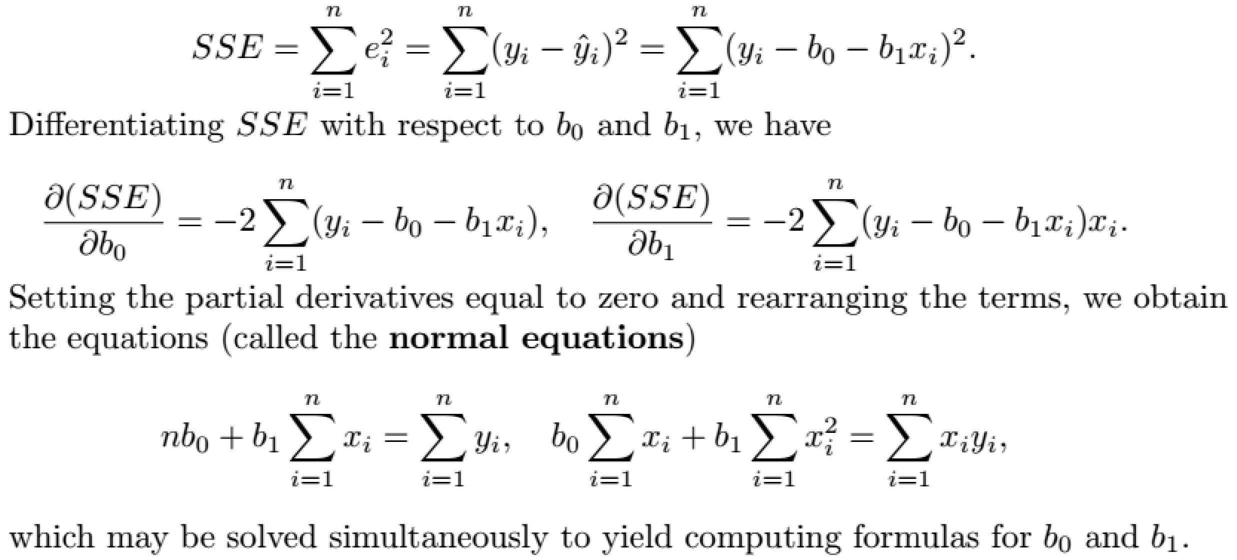
Determining Regression Equation:

- There are several methods for estimating the regression parameters, here we will use Method of Least Sq. to estimate the parameters.
- The formula for parameters estimation by least square is given below:

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

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

- We shall find b0 and b1, so that the sum of the squares of the residuals is a minimum.
- The residual sum of squares is often called the sum of squares of the errors about the regression line and is denoted by SSE.
- This minimization procedure for estimating the parameters is called the **method of least squares**.



$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Differentiating SSE with respect to b_0 and b_1, we have

$$\frac{\partial(SSE)}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i), \quad \frac{\partial(SSE)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i.$$

Setting the partial derivatives equal to zero and rearranging the terms, we obtain the equations (called the **normal equations**)

$$nb_0 + b_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

which may be solved simultaneously to yield computing formulas for b_0 and b_1.

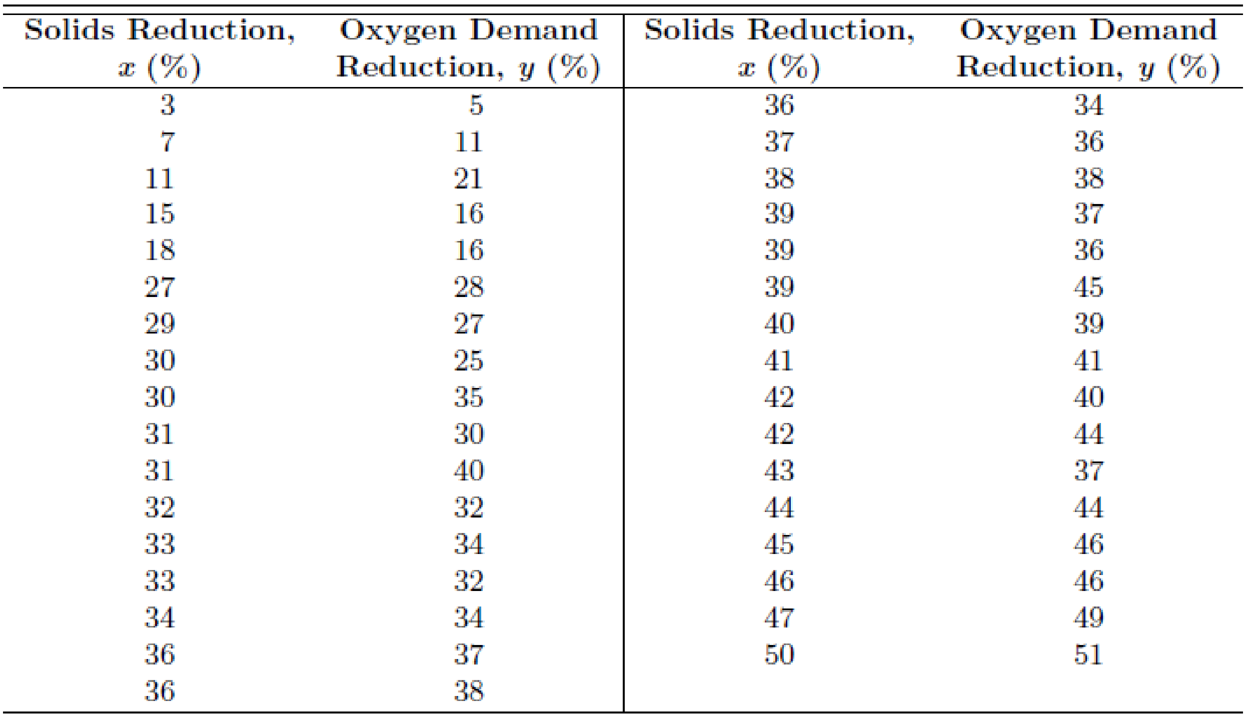
Example:

Consider the experimental data in Table 11.1, which were obtained from 33 samples of chemically treated waste in a study conducted at Virginia Tech. Readings on x, the percent reduction in total solids, and y, the percent reduction in chemical oxygen demand, were recorded.

The data of Table 11.1 are plotted in a scatter diagram in Figure 11.3. From an inspection of this scatter diagram, it can be seen that the points closely follow a straight line, indicating that the assumption of linearity between the two variables appears to be reasonable.

Solution:

Table 11.1: Measures of Reduction in Solids and Oxygen Demand



Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)	Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

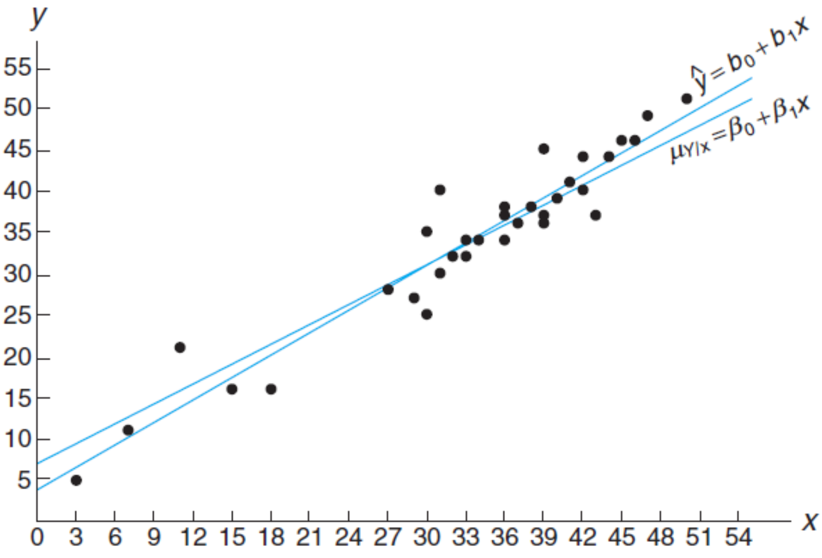


Figure 11.3: Scatter diagram with regression lines.

The fitted regression line and a hypothetical true regression line are shown on the scatter diagram of Figure 11.3.

Estimate the regression line for the pollution data of Table 11.1.

$$\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41{,}355, \quad \sum_{i=1}^{33} x_i^2 = 41{,}086$$

Therefore,

$$b_1 = \frac{(33)(41{,}355) - (1104)(1124)}{(33)(41{,}086) - (1104)^2} = 0.903643 \text{ and}$$

$$b_0 = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Thus, the estimated regression line is given by

$$\hat{y} = 3.8296 + 0.9036x.$$

Correlation Coefficient

- To determine the strength of the 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), named after statistician Karl Pearson, who pioneered the research in this area.
- The correlation coefficient computed from the sample data measures the strength and direction of a linear relationship between two variables. The symbol for the sample correlation coefficient is r. The symbol for the population correlation coefficient is ρ (Greek letter rho).

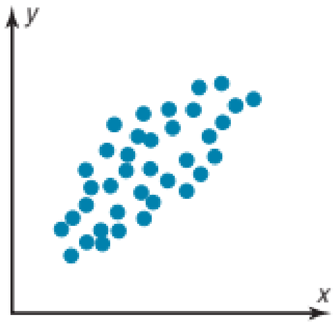
Properties of Correlation Coefficient:

- Correlation coefficient is symmetric $r_{xy} = r_{yx}$
- Correlation coefficient does not depend on units.
- The correlation coefficient lies between – 1 to + 1 i.e. – 1 ≤ r ≤ +1

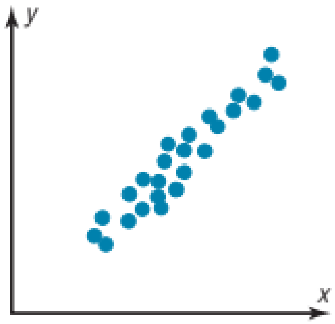
$$r = \frac{n(\Sigma\, xy) - (\Sigma\, x)(\Sigma\, y)}{\sqrt{[n(\Sigma\, x^2) - (\Sigma\, x)^2][n(\Sigma\, y^2) - (\Sigma\, y)^2]}}$$

where n is the number of data pairs.

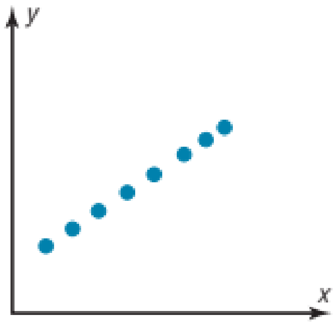
- Since the value of r is computed from data obtained from samples, there are two possibilities when r is not equal to zero: either the value of r is high enough to conclude that there is a significant linear relationship between the variables, or the value of r is due to chance.



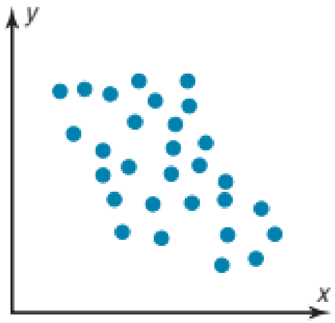
(a) $r = 0.50$



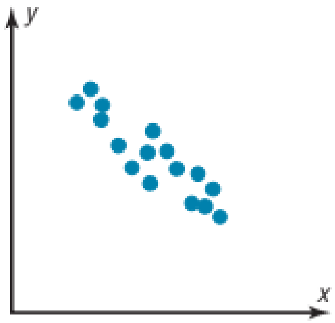
(b) $r = 0.90$



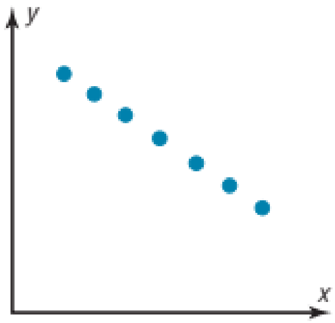
(c) $r = 1.00$



(d) $r = -0.50$



(e) $r = -0.90$



(f) $r = -1.00$

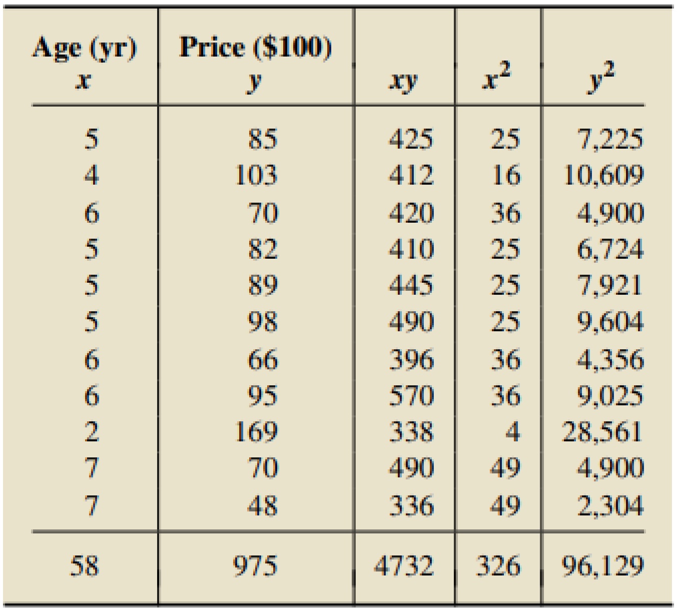
Example:

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 14.10 on the next page.

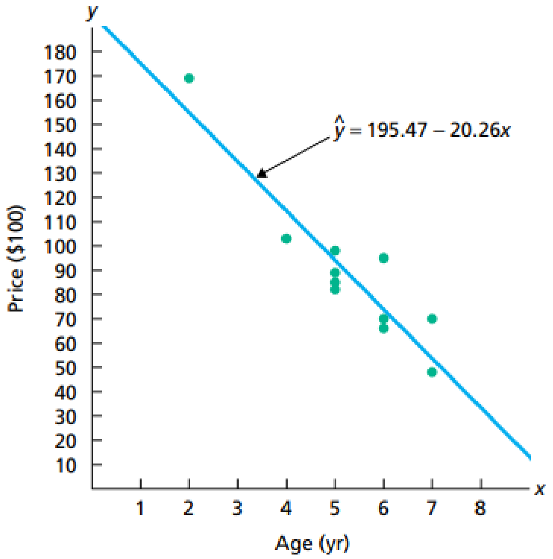
a. Compute the linear correlation coefficient, r, of the data.

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.



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



Solution:

a. We apply Formula 14.3 to find the linear correlation coefficient. To accomplish that, we need a table of values for x, y, xy, x^2, y^2, and their sums, as shown in Table 14.10. Referring to the last row of Table 14.10, we get

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

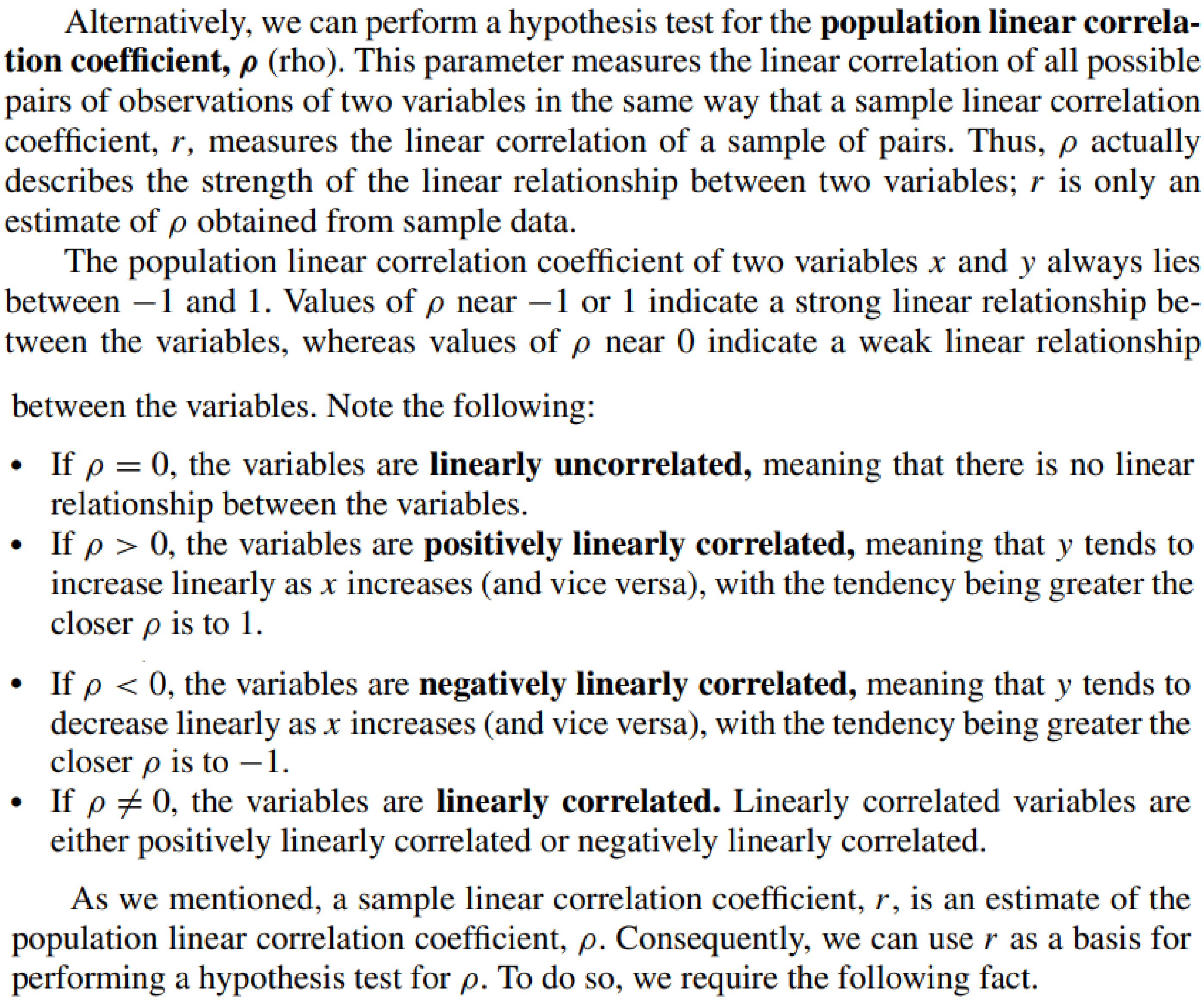
$$= \frac{4732 - (58)(975)/11}{\sqrt{[326 - (58)^2/11][96{,}129 - (975)^2/11]}} = -0.924.$$

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. Figure 14.16 on page 663 shows that to be the case.

Inferences in Correlation

Frequently, we want to decide whether two variables are linearly correlated, that is, whether there is a linear relationship between the two variables. In the context of regression, we can make that decision by performing a hypothesis test for the slope of the population regression line

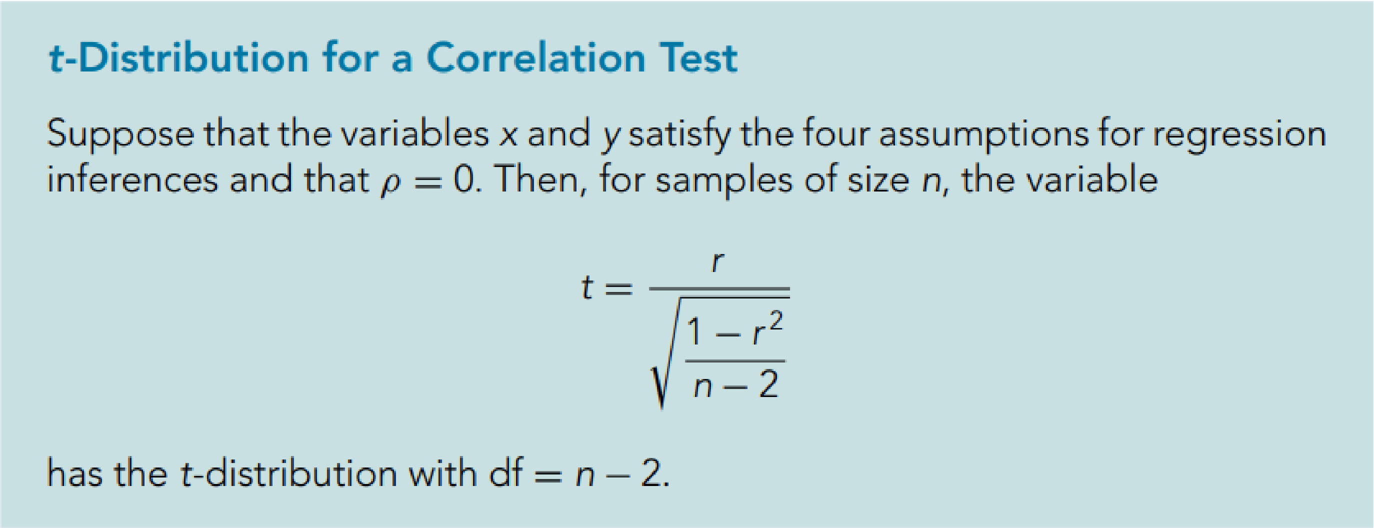


Alternatively, we can perform a hypothesis test for the **population linear correlation coefficient, ρ** (rho). This parameter measures the linear correlation of all possible pairs of observations of two variables in the same way that a sample linear correlation coefficient, r, measures the linear correlation of a sample of pairs. Thus, ρ actually describes the strength of the linear relationship between two variables; r is only an estimate of ρ obtained from sample data.

The population linear correlation coefficient of two variables x and y always lies between -1 and 1. Values of ρ near -1 or 1 indicate a strong linear relationship between the variables, whereas values of ρ near 0 indicate a weak linear relationship between the variables. Note the following:

- If $\rho = 0$, the variables are **linearly uncorrelated,** meaning that there is no linear relationship between the variables.
- If $\rho > 0$, the variables are **positively linearly correlated,** meaning that y tends to increase linearly as x increases (and vice versa), with the tendency being greater the closer ρ is to 1.
- If $\rho < 0$, the variables are **negatively linearly correlated,** meaning that y tends to decrease linearly as x increases (and vice versa), with the tendency being greater the closer ρ is to -1.
- If $\rho \neq 0$, the variables are **linearly correlated.** Linearly correlated variables are either positively linearly correlated or negatively linearly correlated.

As we mentioned, a sample linear correlation coefficient, r, is an estimate of the population linear correlation coefficient, ρ. Consequently, we can use r as a basis for performing a hypothesis test for ρ. To do so, we require the following fact.



t-Distribution for a Correlation Test

Suppose that the variables x and y satisfy the four assumptions for regression inferences and that $\rho = 0$. Then, for samples of size n, the variable

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

has the t-distribution with $\mathrm{df} = n - 2$.

PROCEDURE:

Purpose To perform a hypothesis test for a population linear correlation coefficient, ρ

Assumptions The four assumptions for regression inferences

Step 1 The null hypothesis is H_0: $\rho = 0$, and the alternative hypothesis is

$$\underset{\text{(Two tailed)}}{H_a\colon \rho \neq 0} \quad \text{or} \quad \underset{\text{(Left tailed)}}{H_a\colon \rho < 0} \quad \text{or} \quad \underset{\text{(Right tailed)}}{H_a\colon \rho > 0}.$$

Step 2 Decide on the significance level, α.

Step 3 Compute the value of the test statistic

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

and denote that value t_0.

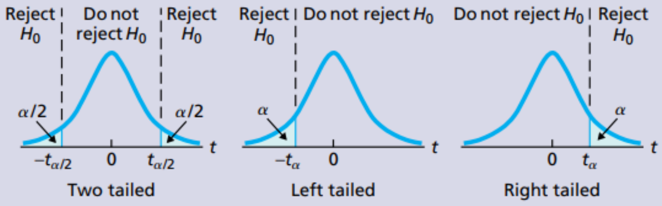
CRITICAL-VALUE APPROACH **OR** **P-VALUE APPROACH**

CRITICAL-VALUE APPROACH

Step 4 The critical value(s) are

$$\underset{\text{(Two tailed)}}{\pm t_{\alpha/2}} \quad \text{or} \quad \underset{\text{(Left tailed)}}{-t_{\alpha}} \quad \text{or} \quad \underset{\text{(Right tailed)}}{t_{\alpha}}$$

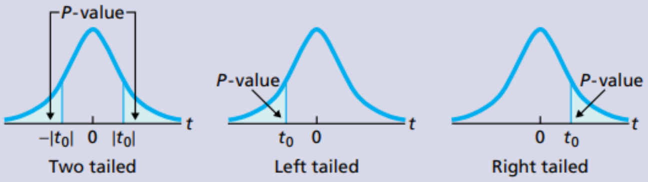
with df $= n - 2$. Use Table IV to find the critical value(s).



Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

P-VALUE APPROACH

Step 4 The t-statistic has df $= n - 2$. Use Table IV to estimate the P-value, or obtain it exactly by using technology.



Step 5 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

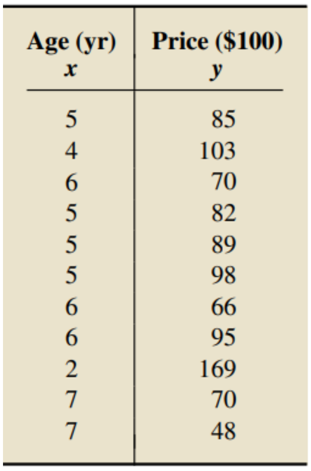
Step 6 Interpret the results of the hypothesis test.

Example:

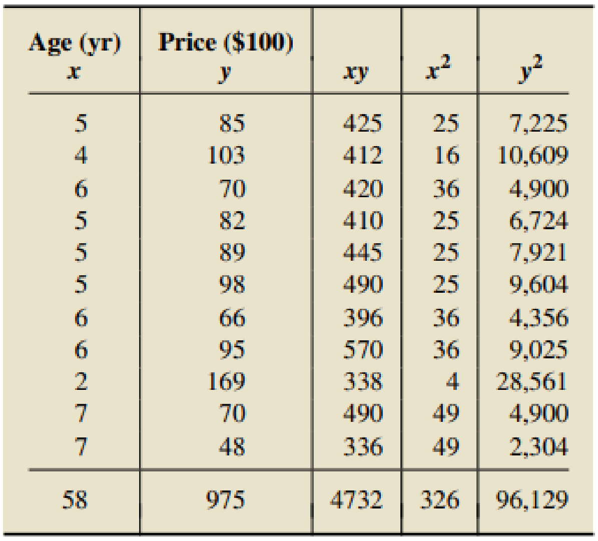
Age and Price of Orions The data on age and price for a sample of 11 Orions are repeated in Table 15.5. At the 5% significance level, do the data provide sufficient evidence to conclude that age and price of Orions are negatively linearly correlated?

TABLE 15.5

Age and price data for a sample of 11 Orions



Age (yr) x	Price ($100) y
5	85
4	103
6	70
5	82
5	89
5	98
6	66
6	95
2	169
7	70
7	48



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

Solution:

Step 1 State the null and alternative hypotheses.

Let ρ denote the population linear correlation coefficient for the variables age and price of Orions. Then the null and alternative hypotheses are, respectively,

H_0: $\rho = 0$ (age and price are linearly uncorrelated)

H_a: $\rho < 0$ (age and price are negatively linearly correlated).

Note that the hypothesis test is left tailed.

Step 2 Decide on the significance level, α.

We are to use $\alpha = 0.05$.

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

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

Step 3 Compute the value of the test statistic

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

In Example 14.13 on page 670, we found that $r = -0.924$, so the value of the test statistic is

$$t = \frac{-0.924}{\sqrt{\dfrac{1-(-0.924)^2}{11-2}}} = -7.249.$$

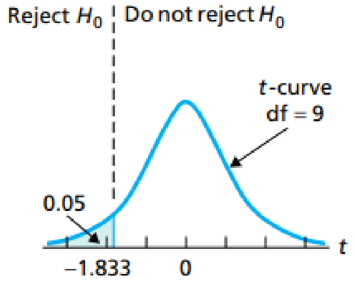
CRITICAL-VALUE APPROACH

OR

Step 4 The critical value for a left-tailed test is $-t_\alpha$ with df $= n - 2$. Use Table IV to find the critical value.

For $n = 11$, df $= 9$. Also, $\alpha = 0.05$. From Table IV, for df $= 9$, $t_{0.05} = 1.833$. Consequently, the critical value is $-t_{0.05} = -1.833$, as shown in Fig. 15.12A.

FIGURE 15.12A



Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

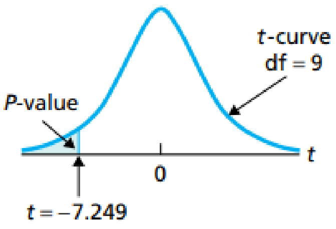
The value of the test statistic, found in Step 3, is $t = -7.249$. Figure 15.12A shows that this value falls in the rejection region, so we reject H_0. The test results are statistically significant at the 5% level.

P-VALUE APPROACH

Step 4 The t-statistic has df $= n - 2$. Use Table IV to estimate the P-value or obtain it exactly by using technology.

From Step 3, the value of the test statistic is $t = -7.249$. Because the test is left tailed, the P-value is the probability of observing a value of t of -7.249 or less if the null hypothesis is true. That probability equals the shaded area shown in Fig. 15.12B.

FIGURE 15.12B



For $n = 11$, df $= 9$. Referring now to Fig. 15.12B and Table IV, we find that $P < 0.005$. (Using technology, we obtain $P = 0.0000244$.)

Step 5 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

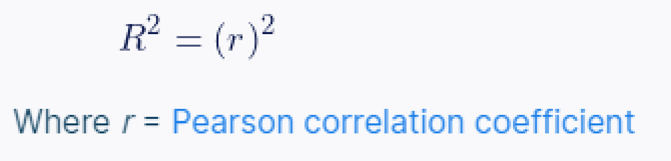
From Step 4, $P < 0.005$. Because the P-value is less than the specified significance level of 0.05, we reject H_0. The test results are statistically significant at the 5% level and (see Table 9.8 on page 408) provide very strong evidence against the null hypothesis.

Step 6 Interpret the results of the hypothesis test.

Interpretation At the 5% significance level, the data provide sufficient evidence to conclude that age and price of Orions are negatively linearly correlated. Prices for 2- to 7-year-old Orions tend to decrease linearly with increasing age. ■

A Measure of Quality of Fit: Coefficient of Determination

This quantity is a measure of the proportion of variability explained by the fitted model.



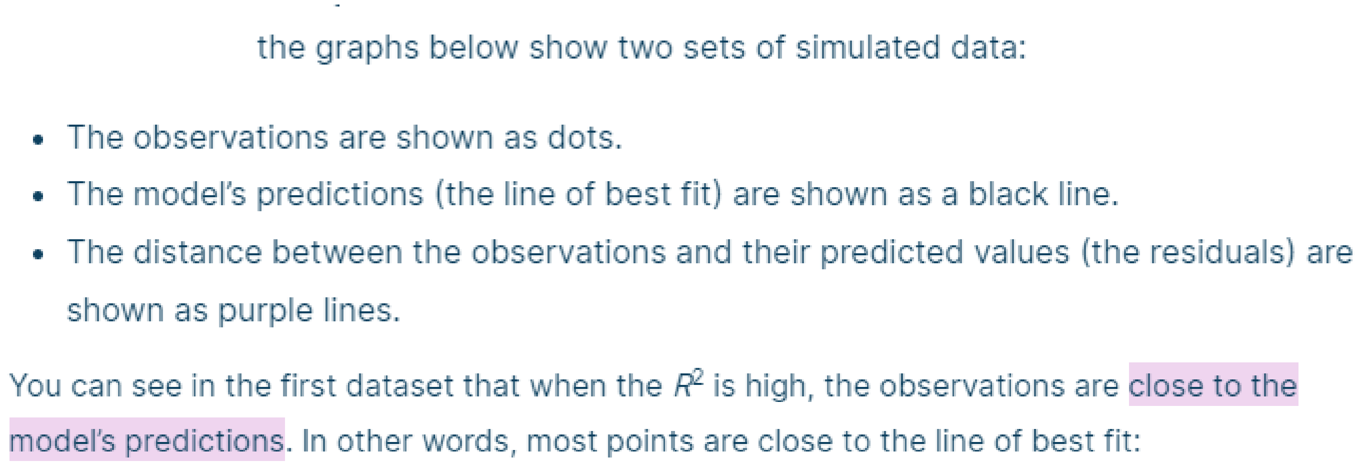
$$R^2 = (r)^2$$

Where r = Pearson correlation coefficient

Interpreting the coefficient of determination:

- If the R^2 is 0, the linear regression model doesn't allow you to predict exam scores any better than simply estimating that everyone has an average exam score.
- If the R^2 is between 0 and 1, the model allows you to partially predict exam scores. The model's estimates are not perfect, but they're better than simply using the average exam score.
- If the R^2 is 1, the model allows you to perfectly predict anyone's exam score.

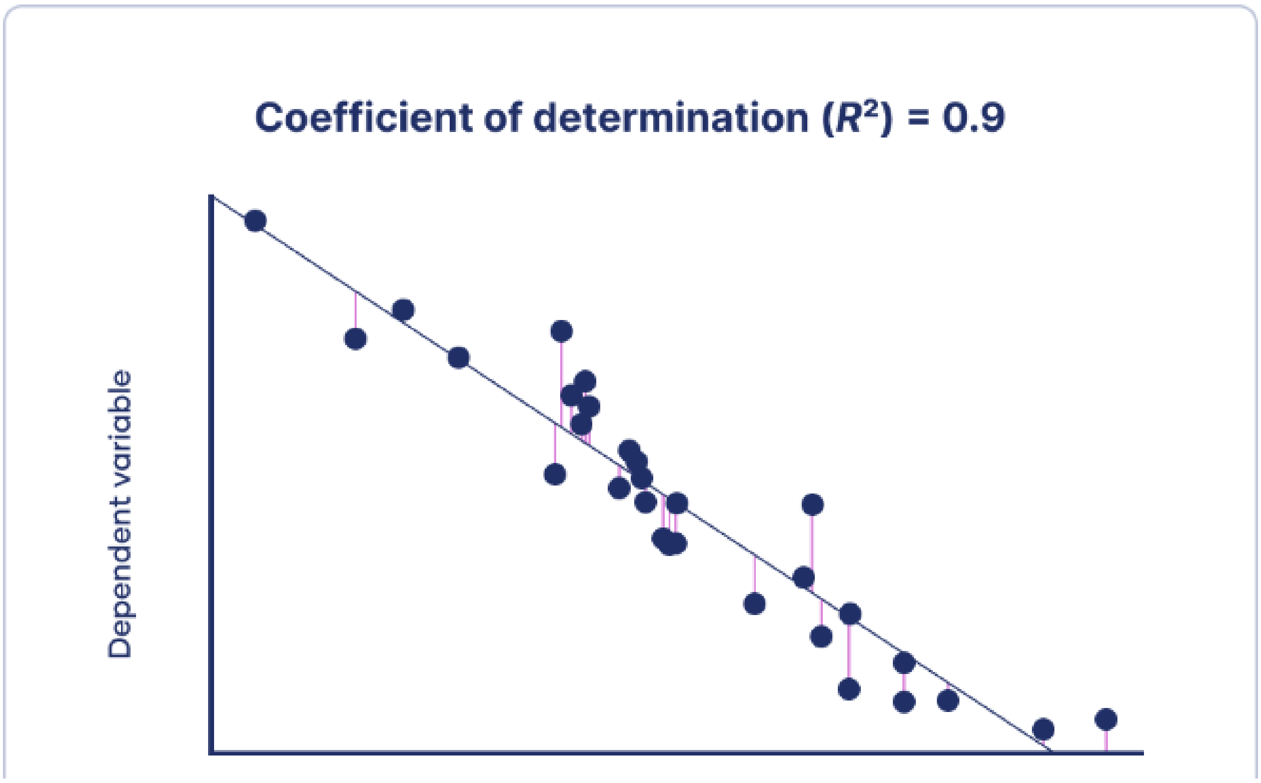
Example:



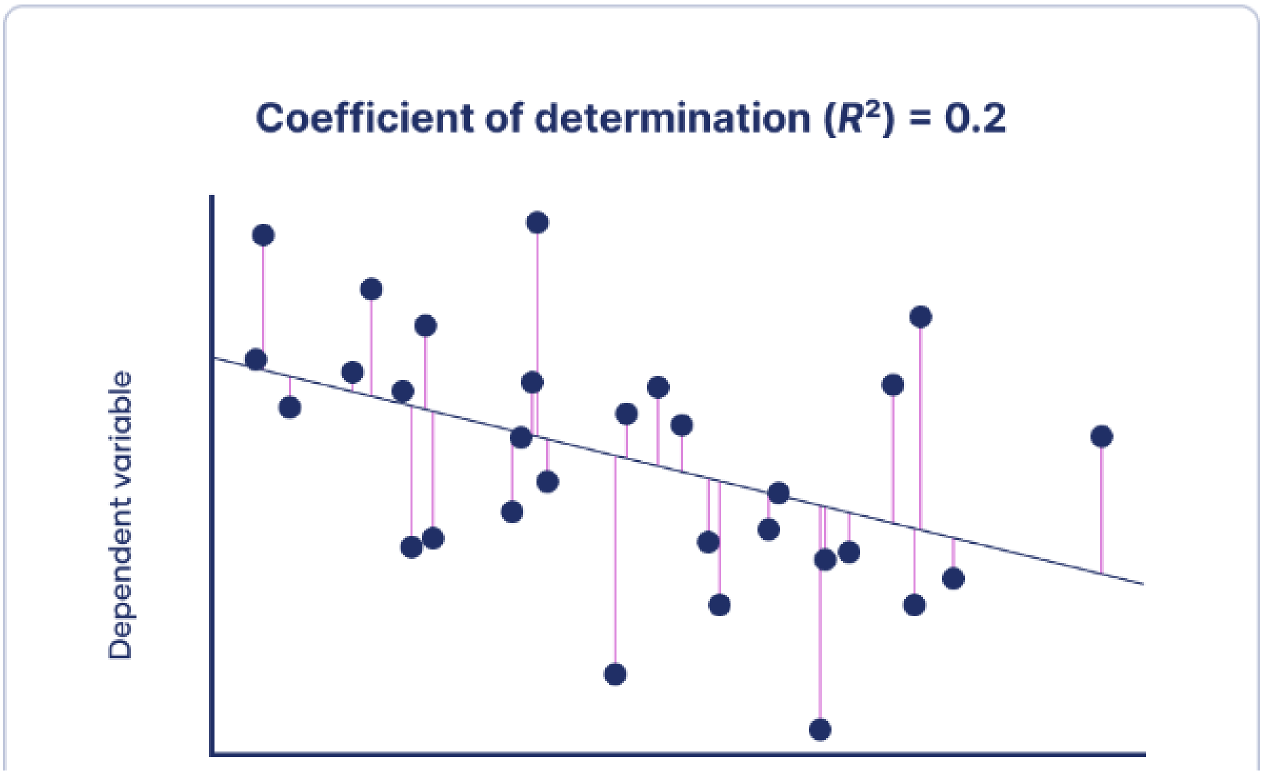
the graphs below show two sets of simulated data:

- The observations are shown as dots.
- The model's predictions (the line of best fit) are shown as a black line.
- The distance between the observations and their predicted values (the residuals) are shown as purple lines.

You can see in the first dataset that when the R^2 is high, the observations are close to the model's predictions. In other words, most points are close to the line of best fit:



In contrast, you can see in the second dataset that when the R^2 is low, the observations are far from the model's predictions. In other words, when the R^2 is low, many points are far from the line of best fit:



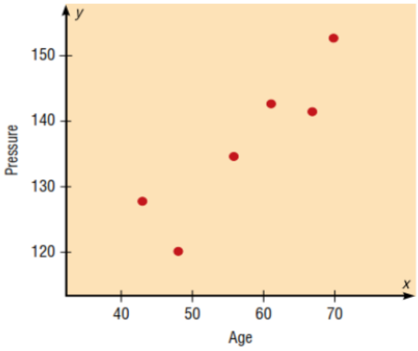
Example:

Construct a scatter plot for the data obtained in a study of age and systolic blood pressure of six randomly selected subjects. The data are shown in the following table.
Using the equation of the regression line, predict the blood pressure for a person who is 50 years old.

Subject	Age, x	Pressure, y
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152

Solution:

Scatter plot



Subject	Age, x	Pressure, y	xy	x^2	y^2
A	43	128	5,504	1,849	16,384
B	48	120	5,760	2,304	14,400
C	56	135	7,560	3,136	18,225
D	61	143	8,723	3,721	20,449
E	67	141	9,447	4,489	19,881
F	70	152	10,640	4,900	23,104
	$\Sigma x = 345$	$\Sigma y = 819$	$\Sigma xy = 47{,}634$	$\Sigma x^2 = 20{,}399$	$\Sigma y^2 = 112{,}443$

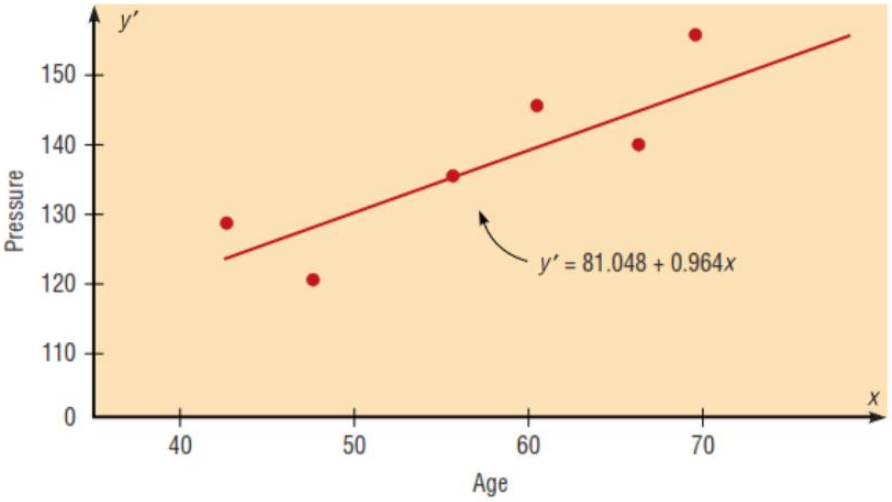
Regression coefficients and regression equation

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(819)(20{,}399) - (345)(47{,}634)}{(6)(20{,}399) - (345)^2} = 81.048$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(6)(47{,}634) - (345)(819)}{(6)(20{,}399) - (345)^2} = 0.964$$

$$y' = 81.048 + 0.964x$$

Regression line



Prediction Using Regression Equation

Substituting 50 for x in the regression line $y' = 81.048 + 0.964x$ gives

$$y' = 81.048 + (0.964)(50) = 129.248 \text{ (rounded to 129)}$$

In other words, the predicted systolic blood pressure for a 50-year-old person is 129.

Correlation coefficient

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

where n is the number of data pairs.

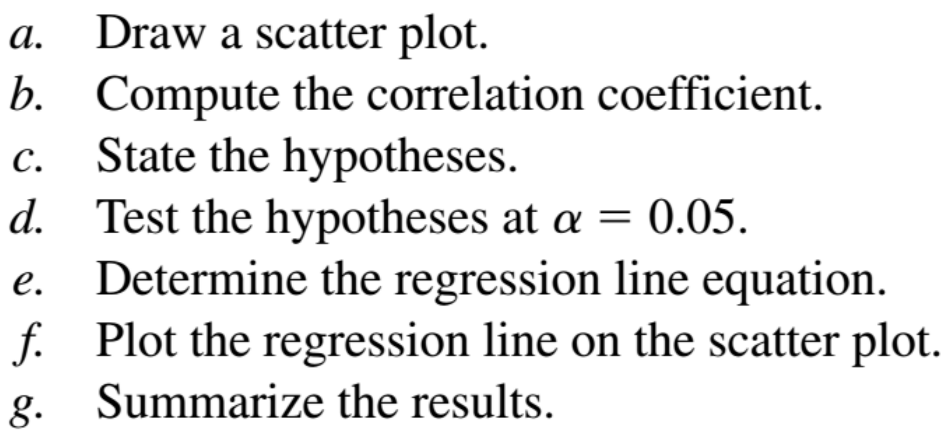
$$r = 0.897$$

Coefficient of Determination

$$R = (0.897)^2 = 0.804609$$

Problem Set

In Question 1 and 2, Do a complete regression analysis by performing the following steps.



a. Draw a scatter plot.

b. Compute the correlation coefficient.

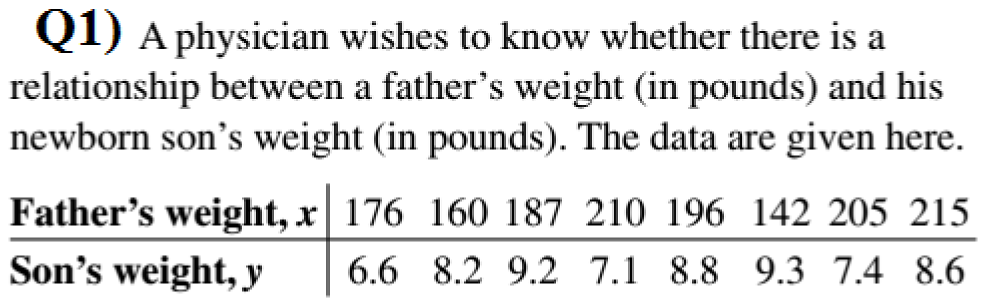
c. State the hypotheses.

d. Test the hypotheses at $\alpha = 0.05$.

e. Determine the regression line equation.

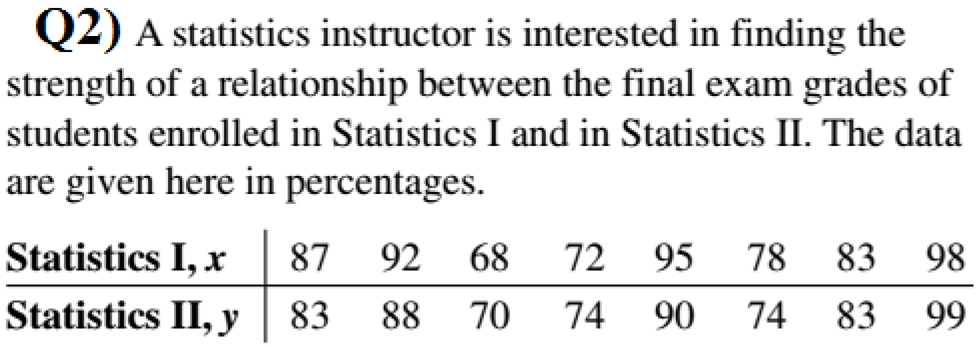
f. Plot the regression line on the scatter plot.

g. Summarize the results.



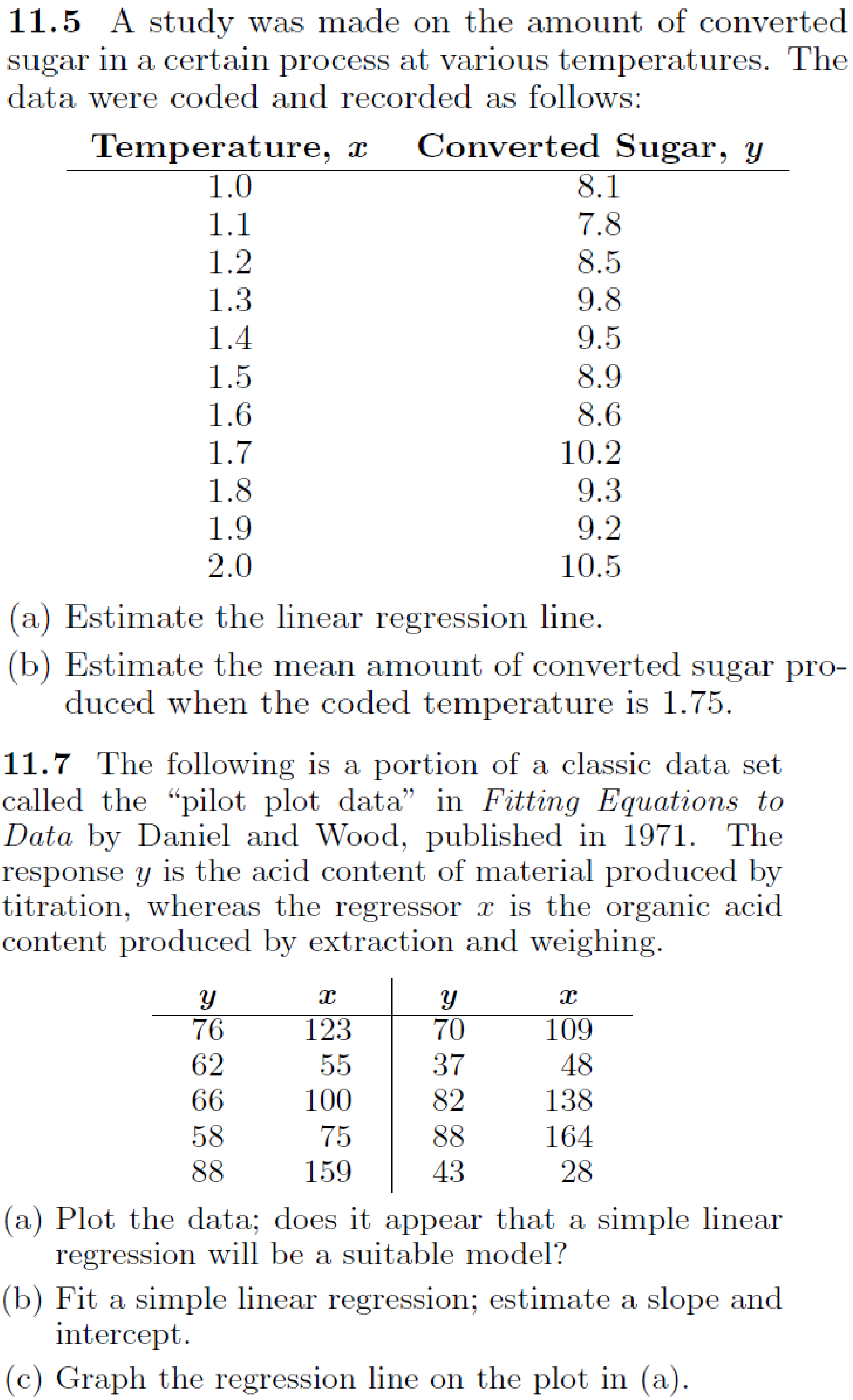
Q1) A physician wishes to know whether there is a relationship between a father's weight (in pounds) and his newborn son's weight (in pounds). The data are given here.

Father's weight, *x*	176	160	187	210	196	142	205	215
Son's weight, *y*	6.6	8.2	9.2	7.1	8.8	9.3	7.4	8.6



Q2) A statistics instructor is interested in finding the strength of a relationship between the final exam grades of students enrolled in Statistics I and in Statistics II. The data are given here in percentages.

Statistics I, *x*	87	92	68	72	95	78	83	98
Statistics II, *y*	83	88	70	74	90	74	83	99



11.5 A study was made on the amount of converted sugar in a certain process at various temperatures. The data were coded and recorded as follows:

Temperature, x	**Converted Sugar, y**
1.0	8.1
1.1	7.8
1.2	8.5
1.3	9.8
1.4	9.5
1.5	8.9
1.6	8.6
1.7	10.2
1.8	9.3
1.9	9.2
2.0	10.5

(a) Estimate the linear regression line.

(b) Estimate the mean amount of converted sugar produced when the coded temperature is 1.75.

11.7 The following is a portion of a classic data set called the "pilot plot data" in *Fitting Equations to Data* by Daniel and Wood, published in 1971. The response y is the acid content of material produced by titration, whereas the regressor x is the organic acid content produced by extraction and weighing.

y	x	y	x
76	123	70	109
62	55	37	48
66	100	82	138
58	75	88	164
88	159	43	28

(a) Plot the data; does it appear that a simple linear regression will be a suitable model?

(b) Fit a simple linear regression; estimate a slope and intercept.

(c) Graph the regression line on the plot in (a).

11.8 A mathematics placement test is given to all entering freshmen at a small college. A student who receives a grade below 35 is denied admission to the regular mathematics course and placed in a remedial class. The placement test scores and the final grades for 20 students who took the regular course were recorded.

(a) Plot a scatter diagram.

(b) Find the equation of the regression line to predict course grades from placement test scores.

(c) Graph the line on the scatter diagram.

(d) If 60 is the minimum passing grade, below which placement test score should students in the future be denied admission to this course?

Placement Test	Course Grade
50	53
35	41
35	61
40	56
55	68
65	36
35	11
60	70
90	79
35	59
90	54
80	91
60	48
60	71
60	71
40	47
55	53
50	68
65	57
50	79