Simple Linear Regression and Correlation

From Walpole (chap # 11)

The Simple Linear Regression (SLR) Model

Simple Linear Regression Model	$$Y = \beta_0 + \beta_1 x + \epsilon.$$

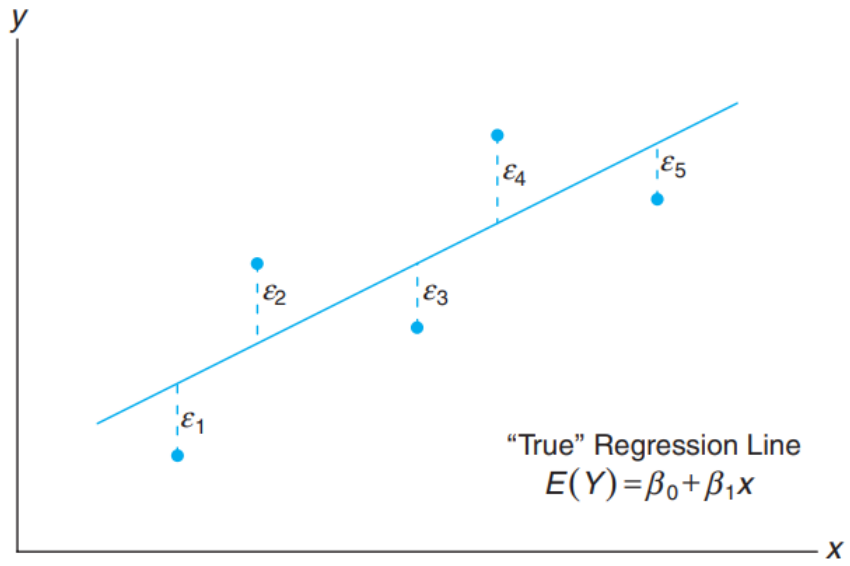


Figure 11.2: Hypothetical (x, y) data scattered around the true regression line for $n = 5$.

The Fitted Regression Line

An important aspect of regression analysis is, very simply, to estimate the parameters β_0 and β_1 (i.e., estimate the so-called **regression coefficients**). The method of estimation will be discussed in the next section. Suppose we denote the estimates b_0 for β_0 and b_1 for β_1. Then the estimated or **fitted regression** line is given by

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

where $\hat{y}$ is the predicted or fitted value. Obviously, the fitted line is an estimate of the true regression line. We expect that the fitted line should be closer to the true regression line when a large amount of data are available. In the following example, we illustrate the fitted line for a real-life pollution study.

One of the more challenging problems confronting the water pollution control field is presented by the tanning industry. Tannery wastes are chemically complex. They are characterized by high values of chemical oxygen demand, volatile solids, and other pollution measures. 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.

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		

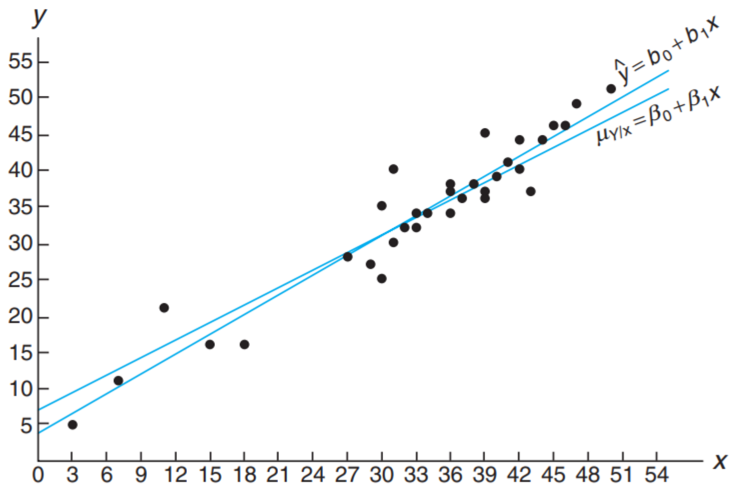
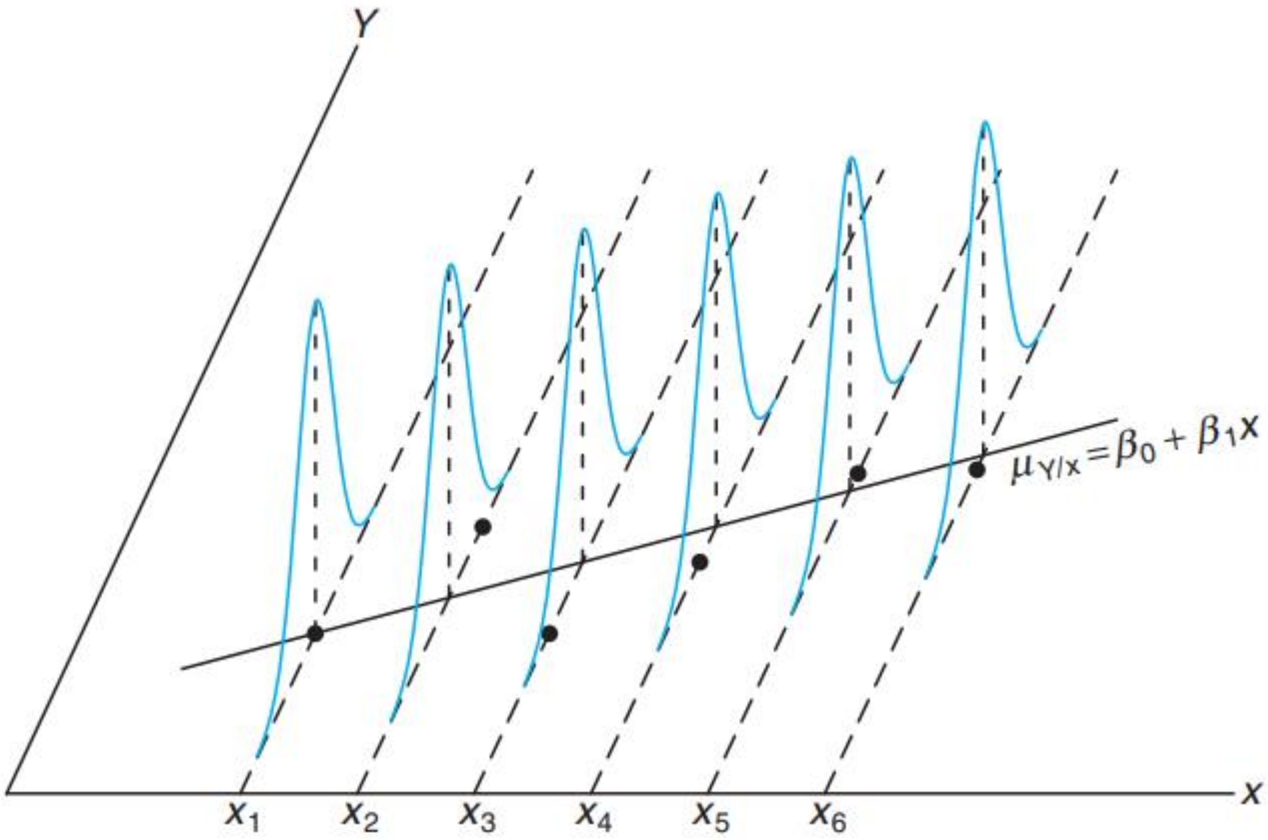


Figure 11.3: Scatter diagram with regression lines.

The fitted regression line and a *hypothetical true regression line* are shown on the scatter diagram of Figure 11.3. This example will be revisited as we move on to the method of estimation, discussed in Section 11.3.

Another Look at the Model Assumptions



11.3 Least Squares and the Fitted Model

Residual: Error in Fit Given a set of regression data $\{(x_i, y_i); i = 1, 2, \dots, n\}$ and a fitted model, $\hat{y}_i = b_0 + b_1 x_i$, the ith residual e_i is given by

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n.$$

Estimating the Regression Coefficients Given the sample $\{(x_i, y_i);\ i = 1, 2, \dots, n\}$, the least squares estimates b_0 and b_1 of the regression coefficients β_0 and β_1 are computed from the formulas

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}.$$

Exercise questions (11.1 to 11.13) Page # 398 to 400

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.

(c) Plot the residuals versus temperature. Comment.

11.6 In a certain type of metal test specimen, the normal stress on a specimen is known to be functionally related to the shear resistance. The following is a set of coded experimental data on the two variables:

Normal Stress, x	Shear Resistance, y
26.8	26.5
25.4	27.3
28.9	24.2
23.6	27.1
27.7	23.6
23.9	25.9
24.7	26.3
28.1	22.5
26.9	21.7
27.4	21.4
22.6	25.8
25.6	24.9

(a) Estimate the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$.

(b) Estimate the shear resistance for a normal stress of 24.5.

	x	y	xy	x^2	y^2
	26.8	**26.5**	**710.2**	**718.24**	**702.25**
	25.4	**27.3**	**693.42**	**645.16**	**745.29**
	28.9	**24.2**	**699.38**	**835.21**	**585.64**
	23.6	**27.1**	**639.56**	**556.96**	**734.41**
	27.7	**23.6**	**653.72**	**767.29**	**556.96**
	23.9	**25.9**	**619.01**	**571.21**	**670.81**
	24.7	**26.3**	**649.61**	**610.09**	**691.69**
	28.1	**22.5**	**632.25**	**789.61**	**506.25**
	26.9	**21.7**	**583.73**	**723.61**	**470.89**
	27.4	**21.4**	**586.36**	**750.76**	**457.96**
	22.6	**25.8**	**583.08**	**510.76**	**665.64**
	25.6	**24.9**	**637.44**	**655.36**	**620.01**
Total	311.6	297.2	7687.76	8134.26	7407.8

$$b_1 = \frac{(12(7687) - (311.6)(297.2))}{(12(8134.26) - 311.6^2))} = -0.6860$$

$$b_0 = \frac{297.2}{12} - \frac{(-0.6860)(311.6)}{12} = 42.5818$$

$\hat{y} = 42.5818 - 0.6860\, x$ (Estimated Regression Line)

$$Part\ (b)\ \hat{\mathbf{y}} = \mathbf{42.5818 - 0.6860\ (24.5) = 25.7748}$$

Gradient Descent Algorithm for computing co-efficients of regression

Example will share separately

Important Formulas

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

OR

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}, \qquad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}, \qquad S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

An unbiased estimate of σ^2 is

$$s^2 = \frac{SSE}{n-2} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}.$$

11.5 Inferences Concerning the Regression Coefficients

Confidence Interval for β_0 A $100(1-\alpha)\%$ confidence interval for the parameter β_0 in the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$ is

$$b_0 - t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2} < \beta_0 < b_0 + t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

Statistical Inference on the Intercept

Confidence intervals and hypothesis testing on the coefficient β_0 may be established from the fact that B_0 is also normally distributed. It is not difficult to show that

$$T = \frac{B_0 - \beta_0}{S\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}$$

Confidence Interval for β_1 A $100(1-\alpha)\%$ confidence interval for the parameter β_1 in the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$ is

$$b_1 - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

T-Statistic for $\boldsymbol{\beta_1}$

$$T = \frac{(B_1 - \beta_1)/(\sigma/\sqrt{S_{xx}})}{S/\sigma} = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

has a t-distribution with $n-2$ degrees of freedom. The statistic T can be used to construct a $100(1-\alpha)\%$ confidence interval for the coefficient β_1.

Practice Problems

11.17 With reference to Exercise 11.5 on page 398,
(a) evaluate s^2;
(b) construct a 95% confidence interval for β_0;
(c) construct a 95% confidence interval for β_1.

11.18 With reference to Exercise 11.6 on page 399,
(a) evaluate s^2;
(b) construct a 99% confidence interval for β_0;
(c) construct a 99% confidence interval for β_1.

11.19 With reference to Exercise 11.3 on page 398,
(a) evaluate s^2;
(b) construct a 99% confidence interval for β_0;
(c) construct a 99% confidence interval for β_1.

11.20 Test the hypothesis that $\beta_0 = 10$ in Exercise 11.8 on page 399 against the alternative that $\beta_0 < 10$. Use a 0.05 level of significance.

11.21 Test the hypothesis that $\beta_1 = 6$ in Exercise 11.9 on page 399 against the alternative that $\beta_1 < 6$. Use a 0.025 level of significance.

11.17 With reference to Exercise 11.5 on page 398,

(a) evaluate s^2;

(b) construct a 95% confidence interval for β_0;

(c) construct a 95% confidence interval for β_1.

Solution:

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

Solution:

$$\hat{y} \qquad (y-\hat{y})^2$$

	x	y	xy	x^2	y^2	ycap	(y-ycap)^2
	1	8.1	8.1	1	65.61	8.22272	0.01506
	1.1	7.8	8.58	1.21	60.84	8.40363	0.36437
	1.2	8.5	10.2	1.44	72.25	8.58454	0.00715
	1.3	9.8	12.74	1.69	96.04	8.76545	1.07030
	1.4	9.5	13.3	1.96	90.25	8.94636	0.30652
	1.5	8.9	13.35	2.25	79.21	9.12727	0.05165
	1.6	8.6	13.76	2.56	73.96	9.30817	0.50151
	1.7	10.2	17.34	2.89	104.04	9.48908	0.50540
	1.8	9.3	16.74	3.24	86.49	9.66999	0.13689
	1.9	9.2	17.48	3.61	84.64	9.85090	0.42367
	2	10.5	21	4	110.25	10.03181	0.21920
Sum	16.5	100.4	152.59	25.85	923.58	100.39992	3.60173

$$b_0 = 6.41363 \ and \ b_1 = 1.80909$$

$$\hat{y} = 6.41363 + 1.80909x$$

$$(a)\, s^2 = \frac{SSE}{n-2} = \frac{\sum(y-\hat{y})^2}{n-2} = \frac{3.60173}{9} = 0.40019,$$

$$s = 0.63261$$

Or

$$s^2 = (S_{yy} - b_1 S_{xy})/(\text{n}-2)$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 923.58 - \frac{100.4^2}{11} = 7.2018$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$

$$S_{xy} = 152.59 - 16.5 * \frac{100.4}{11} = 1.99$$

$$s^2 = \frac{7.2018 - 1.80909 * 1.99}{9} = 0.40019$$

$$s = 0.63261$$

For (b)

$$b_0 - t_{\alpha/2} \frac{s}{\sqrt{nS_{xx}}} \sqrt{\sum_{i=1}^{n} x_i^2} < \beta_0 < b_0 + t_{\alpha/2} \frac{s}{\sqrt{nS_{xx}}} \sqrt{\sum_{i=1}^{n} x_i^2},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05, \frac{\alpha}{2} = 0.025$$

$$s = \sqrt{0.40079} = 0.63261$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 25.85 - \frac{16.5^2}{11} = 1.1$$

$$Degree\ of\ freedom = n - 2 = 9$$

$$t_{0.025,9} = 2.262$$

$$b_0 = 6.41363$$

$$6.41363 - 2.262 \frac{0.63261}{\sqrt{11 * 1.1}} \sqrt{25.85} < \beta_0 < 6.41363 + 2.262 \frac{0.63261}{\sqrt{11 * 1.1}} \sqrt{25.85}$$

$$4.324 < \beta_0 < 8.503$$

For (c)

Confidence Interval for β_1 A $100(1-\alpha)\%$ confidence interval for the parameter β_1 in the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$ is

$$b_1 - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

$$1-\alpha = 0.95 \Rightarrow \alpha = 0.05, \frac{\alpha}{2} = 0.025$$

$$s = \sqrt{0.40079} = 0.63261$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 25.85 - \frac{16.5^2}{11} = 1.1$$

$$Degree\ of\ freedom = n-2 = 9$$

$$t_{0.025,9} = 2.262$$

$$b_1 = 1.80909$$

$$1.80909 - 2.262\frac{0.63261}{\sqrt{1.1}} < \beta_1 < 1.80909 + 2.262\frac{0.63261}{\sqrt{1.1}}$$

$$0.446 < \beta_1 < 3.172$$

11.20 Test the hypothesis that $\beta_0 = 10$ in Exercise 11.8 on page 399 against the alternative that $\beta_0 < 10$. Use a 0.05 level of significance.

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

Solution:

	x	y	xy	x^2	y^2
	50	53	2650	2500	2809
	35	41	1435	1225	1681
	35	61	2135	1225	3721
	40	56	2240	1600	3136
	55	68	3740	3025	4624
	65	36	2340	4225	1296
	35	11	385	1225	121
	60	70	4200	3600	4900
	90	79	7110	8100	6241
	35	59	2065	1225	3481
	90	54	4860	8100	2916
	80	91	7280	6400	8281
	60	48	2880	3600	2304
	60	71	4260	3600	5041
	60	71	4260	3600	5041
	40	47	1880	1600	2209
	55	53	2915	3025	2809
	50	68	3400	2500	4624
	65	57	3705	4225	3249
	50	79	3950	2500	6241
Total	1110	1173	67690	67100	74725

Statistical Inference on the Intercept

Confidence intervals and hypothesis testing on the coefficient β_0 may be established from the fact that B_0 is also normally distributed. It is not difficult to show that

$$T = \frac{B_0 - \beta_0}{S\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}$$

$$B_0 = 32.51, \beta_0 = 10$$

$$S^2 = (S_{yy} - b_1 Sxy)/(n-2)$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 74725 - \frac{1173^2}{20} = 5928.55$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 67100 - 1110 * \frac{1173}{20} = 2588.5$$

$$b_1 = 0.47107$$

$$S^2 = \frac{5928.55 - 0.47107 * 2588.5}{18} = 261.62141$$

$$S = 16.17472$$

$$\sum x^2 = 67100$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 67100 - \frac{1110^2}{20} = 5495$$

This α means β_0

11.20 The hypotheses are

$$H_0 : \alpha = 10,$$
$$H_1 : \alpha > 10.$$

$\alpha = 0.05$.
Critical region: $t > 1.734$.
Computations: $S_{xx} = 67,100 - 1110^2/20 = 5495$, $S_{yy} = 74,725 - 1173^2/20 = 5928.55$, $S_{xy} = 67,690 - (1110)(1173)/20 = 2588.5$, $s^2 = \frac{5928.55-(0.4711)(2588.5)}{18} = 261.617$ and then $s = 16.175$. Now

$$t = \frac{32.51 - 10}{16.175\sqrt{67,100/(20)(5495)}} = 1.78.$$

Decision: Reject H_0 and claim $\alpha > 10$.

This α means β_0

11.21 Test the hypothesis that $\beta_1 = 6$ in Exercise 11.9 on page 399 against the alternative that $\beta_1 < 6$. Use a 0.025 level of significance.

Advertising Costs ($)	Sales ($)
40	385
20	400
25	395
20	365
30	475
50	440
40	490
20	420
50	560
40	525
25	480
50	510

	x	y	xy	x^2	y^2
	40	385	15400	1600	148225
	20	400	8000	400	160000
	25	395	9875	625	156025
	20	365	7300	400	133225
	30	475	14250	900	225625
	50	440	22000	2500	193600
	40	490	19600	1600	240100
	20	420	8400	400	176400
	50	560	28000	2500	313600
	40	525	21000	1600	275625
	25	480	12000	625	230400
	30	510	15300	900	260100
Total	390	5445	181125	14050	2512925

$$T = \frac{(B_1 - \beta_1)/(\sigma/\sqrt{S_{xx}})}{S/\sigma} = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

has a t-distribution with $n-2$ degrees of freedom. The statistic T can be used to construct a $100(1-\alpha)\%$ confidence interval for the coefficient β_1.

Remaining part: Do it yourself
Solution is attached in next slide (for cross check)

11.21 The hypotheses are

$$H_0: \beta = 6,$$
$$H_1: \beta < 6.$$

$\alpha = 0.025$.
Critical region: $t = -2.228$.
Computations: $S_{xx} = 15,650 - 410^2/12 = 1641.667$, $S_{yy} = 2,512.925 - 5445^2/12 = 42,256.25$, $S_{xy} = 191,325 - (410)(5445)/12 = 5,287.5$, $s^2 = \frac{42,256.25-(3,221)(5,287.5)}{10} = 2,522.521$ and then $s = 50.225$. Now

$$t = \frac{3.221 - 6}{50.225/\sqrt{1641.667}} = -2.24.$$

Decision: Reject H_0 and claim $\beta < 6$.

Solutions

11.17 $S_{xx} = 25.85 - 16.5^2/11 = 1.1$, $S_{yy} = 923.58 - 100.4^2/11 = 7.2018$, $S_{xy} = 152.59 - (165)(100.4)/11 = 1.99$, $a = 6.4136$ and $b = 1.8091$.

(a) $s^2 = \frac{7.2018-(1.8091)(1.99)}{9} = 0.40$.

(b) Since $s = 0.632$ and $t_{0.025} = 2.262$ for 9 degrees of freedom, then a 95% confidence interval is

$$6.4136 \pm (2.262)(0.632)\sqrt{\frac{25.85}{(11)(1.1)}} = 6.4136 \pm 2.0895,$$

which implies $4.324 < \alpha < 8.503$.

(c) $1.8091 \pm (2.262)(0.632)/\sqrt{1.1}$ implies $0.446 < \beta < 3.172$.

11.18 $S_{xx} = 8134.26 - 311.6^2/12 = 43.0467$, $S_{yy} = 7407.80 - 297.2^2/12 = 47.1467$, $S_{xy} = 7687.76 - (311.6)(297.2)/12 = -29.5333$, $a = 42.5818$ and $b = -0.6861$.

(a) $s^2 = \frac{47.1467-(-0.6861)(-29.5333)}{10} = 2.688$.

(b) Since $s = 1.640$ and $t_{0.005} = 3.169$ for 10 degrees of freedom, then a 99% confidence interval is

$$42.5818 \pm (3.169)(1.640)\sqrt{\frac{8134.26}{(12)(43.0467)}} = 42.5818 \pm 20.6236,$$

which implies $21.958 < \alpha < 63.205$.

Solutions

(c) $-0.6861 \pm (3.169)(1.640)/\sqrt{43.0467}$ implies $-1.478 < \beta < 0.106$.

11.19 $S_{xx} = 37,125 - 675^2/18 = 11,812.5$, $S_{yy} = 17,142 - 488^2/18 = 3911.7778$, $S_{xy} = 25,005 - (675)(488)/18 = 6705$, $a = 5.8254$ and $b = 0.5676$.

(a) $s^2 = \frac{3911.7778-(0.5676)(6705)}{16} = 6.626$.

(b) Since $s = 2.574$ and $t_{0.005} = 2.921$ for 16 degrees of freedom, then a 99% confidence interval is

$$5.8261 \pm (2.921)(2.574)\sqrt{\frac{37,125}{(18)(11,812.5)}} = 5.8261 \pm 3.1417,$$

which implies $2.686 < \alpha < 8.968$.

(c) $0.5676 \pm (2.921)(2.574)/\sqrt{11,812.5}$ implies $0.498 < \beta < 0.637$.

11.20 The hypotheses are

$$H_0 : \alpha = 10,$$
$$H_1 : \alpha > 10.$$

$\alpha = 0.05$.
Critical region: $t > 1.734$.
Computations: $S_{xx} = 67,100 - 1110^2/20 = 5495$, $S_{yy} = 74,725 - 1173^2/20 = 5928.55$, $S_{xy} = 67,690 - (1110)(1173)/20 = 2588.5$, $s^2 = \frac{5928.55-(0.4711)(2588.5)}{18} = 261.617$ and then $s = 16.175$. Now

$$t = \frac{32.51 - 10}{16.175\sqrt{67,100/(20)(5495)}} = 1.78.$$

Decision: Reject H_0 and claim $\alpha > 10$.

11.21 The hypotheses are

$$H_0 : \beta = 6,$$
$$H_1 : \beta < 6.$$

Solutions

$\alpha = 0.025$.
Critical region: $t = -2.228$.
Computations: $S_{xx} = 15,650 - 410^2/12 = 1641.667$, $S_{yy} = 2,512.925 - 5445^2/12 = 42,256.25$, $S_{xy} = 191,325 - (410)(5445)/12 = 5,287.5$, $s^2 = \frac{42,256.25-(3,221)(5,287.5)}{10} = 2,522.521$ and then $s = 50.225$. Now

$$t = \frac{3.221 - 6}{50.225/\sqrt{1641.667}} = -2.24.$$

Decision: Reject H_0 and claim $\beta < 6$.

Correlation

Correlation Coefficient The measure ρ of linear association between two variables X and Y is estimated by the **sample correlation coefficient** r, where

$$r = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

or

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

- **Population correlation co-efficient is denoted by ρ**
- **r is estimator of ρ**

Co-efficient of determination = r^2 $(gives\ variation\ explained\ by\ response)$

Test statistic for correlation co-efficient

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

which, as before, is a value of the statistic T having a t-distribution with $n-2$ degrees of freedom.

Example 11.11: For the data of Example 11.10, test the hypothesis that there is no linear association among the variables.

Example 11.10: It is important that scientific researchers in the area of forest products be able to study correlation among the anatomy and mechanical properties of trees. For the study *Quantitative Anatomical Characteristics of Plantation Grown Loblolly Pine (Pinus Taeda L.) and Cottonwood (Populus deltoides Bart. Ex Marsh.) and Their Relationships to Mechanical Properties*, conducted by the Department of Forestry and Forest Products at Virginia Tech, 29 loblolly pines were randomly selected for investigation. Table 11.9 shows the resulting data on the specific gravity in grams/cm^3 and the modulus of rupture in kilopascals (kPa). Compute and interpret the sample correlation coefficient.

Table 11.9: Data on 29 Loblolly Pines for Example 11.10

Specific Gravity, x (g/cm^3)	Modulus of Rupture, y (kPa)	Specific Gravity, x (g/cm^3)	Modulus of Rupture, y (kPa)
0.414	29,186	0.581	85,156
0.383	29,266	0.557	69,571
0.399	26,215	0.550	84,160
0.402	30,162	0.531	73,466
0.442	38,867	0.550	78,610
0.422	37,831	0.556	67,657
0.466	44,576	0.523	74,017
0.500	46,097	0.602	87,291
0.514	59,698	0.569	86,836
0.530	67,705	0.544	82,540
0.569	66,088	0.557	81,699
0.558	78,486	0.530	82,096
0.577	89,869	0.547	75,657
0.572	77,369	0.585	80,490
0.548	67,095		

***Solution*:** From the data we find that

$$S_{xx} = 0.11273, \quad S_{yy} = 11{,}807{,}324{,}805, \quad S_{xy} = 34{,}422.27572.$$

Therefore,

$$r = \frac{34{,}422.27572}{\sqrt{(0.11273)(11{,}807{,}324{,}805)}} = 0.9435.$$

Solution:

1. H_0: $\rho = 0$.
2. H_1: $\rho \neq 0$.
3. $\alpha = 0.05$.
4. Critical region: $t < -2.052$ or $t > 2.052$.
5. Computations: $t = \frac{0.9435\sqrt{27}}{\sqrt{1-0.9435^2}} = 14.79$, $P < 0.0001$.
6. Decision: Reject the hypothesis of no linear association.

Exercise questions (11.43 to 11.47) Page # 435 to 436

Class Activity

11.43 Compute and interpret the correlation coefficient for the following grades of 6 students selected at random:

Mathematics grade	70	92	80	74	65	83
English grade	74	84	63	87	78	90

11.46 Test the hypothesis that $\rho = 0$ in Exercise 11.43 against the alternative that $\rho \neq 0$. Use a 0.05 level of significance.

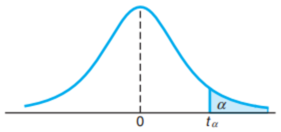


Table A.4 Critical Values of the t-Distribution

	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

SST = SSR + SSE

CONNECTION?

Total variability = Explained variability + Unxplained variability

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

	x	y	xy	x^2	y^2	predicted y	y-ybAR ^ 2	YPRED - YBAR ^2	ei	ei^2
	1	0.1	0.1	1	0.01	0.1041	0.001936	0.00159201	0.0041	1.681E-05
	4	0.12	0.48	16	0.0144	0.1287	0.000576	0.00023409	0.0087	7.569E-05
	6	0.16	0.96	36	0.0256	0.1451	0.000256	1.21E-06	-0.0149	0.00022201
	8	0.18	1.44	64	0.0324	0.1615	0.001296	0.00030625	-0.0185	0.00034225
	10	0.16	1.6	100	0.0256	0.1779	0.000256	0.00114921	0.0179	0.00032041
Total	29	0.72	4.58	217	0.108	0.7173	0.00432	0.00328277	-0.0027	0.00097717

b0	0.0959
b1	0.0082
r	0.879

ymean/ybar	0.144

SST	0.00432
SSR	0.003283
SSE	0.000977
SSR+SSE	0.00426