Section 11.5 Inferences Concerning the Regression Coefficients

Confidence Interval for β, the slope parameter

A (1-α)100% confidence interval for the parameter β in the regression line $\mu_{Y|x} = \alpha + \beta x$ is

$$b - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta < b + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}}$$

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

Example 11.2 Find a 95% confidence interval for β in the regression line $\mu_{Y|x} = \alpha + \beta x$, based on the pollution data of Table 11.1.

Solution From the results given in Example 11.1 we find that

$$S_{xx} = 4152.18, \qquad S_{xy} = 3752.09.$$

In addition, we find that $S_{yy} = 3713.88$. Recall that $b = 0.903643$. Hence,

$$s^2 = \frac{S_{yy} - bS_{xy}}{n-2} = \frac{3713.88 - (0.903643)(3752.09)}{31} = 10.4299.$$

Therefore, taking the square root we obtain $s = 3.2295$. Using Table A.4, we find $t_{0.025} \simeq 2.045$ for 31 degrees of freedom. Therefore, a 95% confidence interval for β is

$$0.903643 - \frac{(2.045)(3.2295)}{\sqrt{4152.18}} < \beta < 0.903643 + \frac{(2.045)(3.2295)}{\sqrt{4152.18}},$$

which simplifies to

$$0.8012 < \beta < 1.0061.$$

Hypothesis testing on the slope parameter, β

To test the null hypothesis H_o:$\beta = \beta_o$ against any suitable alternatives,
use t-distribution with n-2 degrees of freedom to define the critical region, and the following test statistic

$$t = \frac{b - \beta_o}{s/\sqrt{S_{xx}}}$$

Example 11.3 Using the estimated value $b = 0.903643$ of Example 11.1, test the hypothesis that $\beta = 1.0$ against the alternative that $\beta < 1.0$.

Solution

$$H_0: \quad \beta = 1.0,$$
$$H_1: \quad \beta < 1.0,$$
$$t = \frac{0.903643 - 1.0}{3.2295/\sqrt{4152.18}} = -1.92,$$

with $n - 2 = 31$ degrees of freedom ($P \simeq 0.03$),

Decision: The *t*-value is significant at the 0.03 level, suggesting strong evidence that $\beta < 1.0$.

Minitab Example:

One important *t*-test on the slope is the test of the hypothesis

$$H_0: \quad \beta = 0,$$
$$H_1: \quad \beta \neq 0.$$

data of Example 11.1.

$$t = \frac{\text{coefficient}}{\text{standard error}} = \frac{b}{s/\sqrt{s_{xx}}}.$$

Regression Analysis

The regression equation is
COD = 3.83 + 0.904 Per_Red

Predictor	Coef	StDev	T	P
Constant	3.830	1.768	2.17	0.038
Per_Red	0.90364	0.05012	18.03	0.000

S = 3.230 R-Sq = 91.3% R-Sq(adj) = 91.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3390.6	3390.6	325.08	0.000
Error	31	323.3	10.4		
Total	32	3713.9			

Figure 11.6 MINITAB printout for *t*-test for data of Example 11.1.

Statistical Inference on the intercept

Confidence Interval for α, the intercept parameter

A (1-α)100% confidence interval for the parameter α in the regression line $\mu_{Y|x} = \alpha + \beta x$ is

$$a - t_{\alpha/2} s \frac{\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}} < \alpha < a + t_{\alpha/2} s \frac{\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}} \text{ or}$$

$$a - t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} < \alpha < a + t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

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

Example 11.4 Find a 95% confidence interval for α in the regression line $\mu_{Y|x} = \alpha + \beta x$, based on the data of Table 11.1.

Solution In Examples 11.1 and 11.2 we found that

$$S_{xx} = 4152.18 \quad \text{and} \quad s = 3.2295.$$

From Example 11.1 we had

$$\sum_{i=1}^{n} x_i^2 = 41{,}086 \quad \text{and} \quad a = 3.829633.$$

Using Table A.4, we find $t_{0.025} \simeq 2.045$ for 31 degrees of freedom. Therefore, a 95% confidence interval for α is

$$3.829633 - \frac{(2.045)(3.2295)\sqrt{41{,}086}}{\sqrt{(33)(4152.18)}} < \alpha < 3.829633 + \frac{(2.045)(3.2295)\sqrt{41{,}086}}{\sqrt{(33)(4152.18)}},$$

which simplifies to $0.2132 < \alpha < 7.4461$.

Hypothesis testing on the intercept parameter, α

To test the null hypothesis H_o: $\alpha = \alpha_o$ against any suitable alternatives,
use t-distribution with n-2 degrees of freedom to define the critical region, and the following test statistic

$$t = \frac{a - \alpha_o}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

Example 11.5 Using the estimated value $a = 3.829640$ of Example 11.1, test the hypothesis that $\alpha = 0$ at the 0.05 level of significance against the alternative that $\alpha \neq 0$.

Solution

$$H_0: \quad \alpha = 0,$$
$$H_1: \quad \alpha \neq 0,$$
$$t = \frac{3.829633 - 0}{3.2295\sqrt{41{,}086/((33)(4152.18))}} = 2.17$$

df=33-2=31, P-value = $P(|T| > 2.17) = 2P(T > 2.17) \cong 0.038$ from STATISTICA

From Table A.4 directly, P-value < 0.05. So, the null hypothesis of zero intercept is rejected at 0.05 level of significance.

A measure of quality of fit: Coefficient of Determination, R^2

Coefficient of Determination, R^2 = proportion of total variability in the dependent variable Y explained by the fitted model.

$$TSS = SSR + SSE$$

$$\text{Coefficient of Determination, } R^2 = 1 - \frac{SSE}{TSS}$$

$R^2 = 1.0$ if fit is perfect
$R^2 = 0.0$ if fit is poor

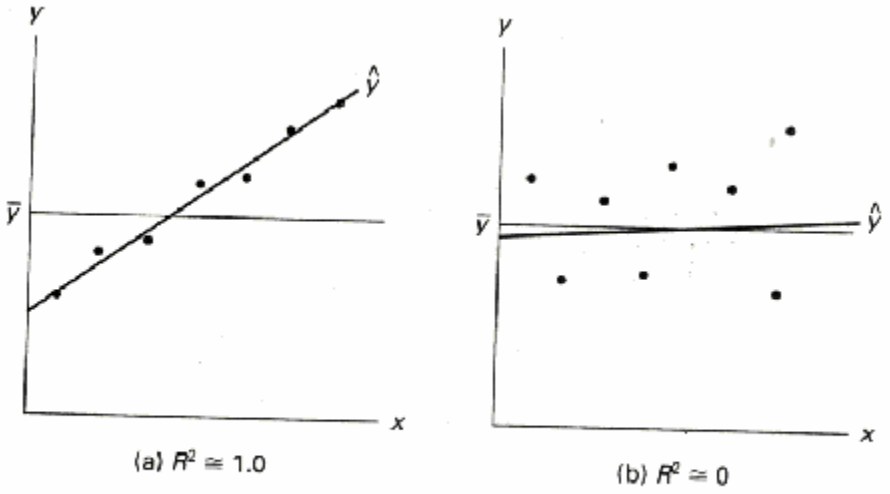


Figure 11.7 Plots depicting a very good fit and a poor fit.

Section 11.6 Prediction

Confidence Interval for $\mu_{Y|x_o}$, the mean response

A (1-α)100% confidence interval for the mean response $\mu_{Y|x_o}$ in the regression line $\mu_{Y|x} = \alpha + \beta x$ is

$$\hat{y}_o - t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_o} < \hat{y}_o + t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}}$$

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

Example 11.6 Using the data of Table 11.1, construct 95% confidence limits for the mean response $\mu_{Y|x}$.

Solution From the regression equation we find for $x_0 = 20\%$ solids reduction, say,

$$\hat{y}_0 = 3.829633 + (0.903643)(20) = 21.9025.$$

In addition, $\bar{x} = 33.4545$, $S_{xx} = 4152.18$, $s = 3.2295$, and $t_{0.025} \simeq 2.045$ for 31 degrees of freedom. Therefore, a 95% confidence interval for $\mu_{Y|20}$ is

$$21.9025 - (2.045)(3.2295)\sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} < \mu_{Y|20}$$

$$< 21.9025 + (2.045)(3.2295)\sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}},$$

or simply $20.1071 < \mu_{Y|20} < 23.6979$.

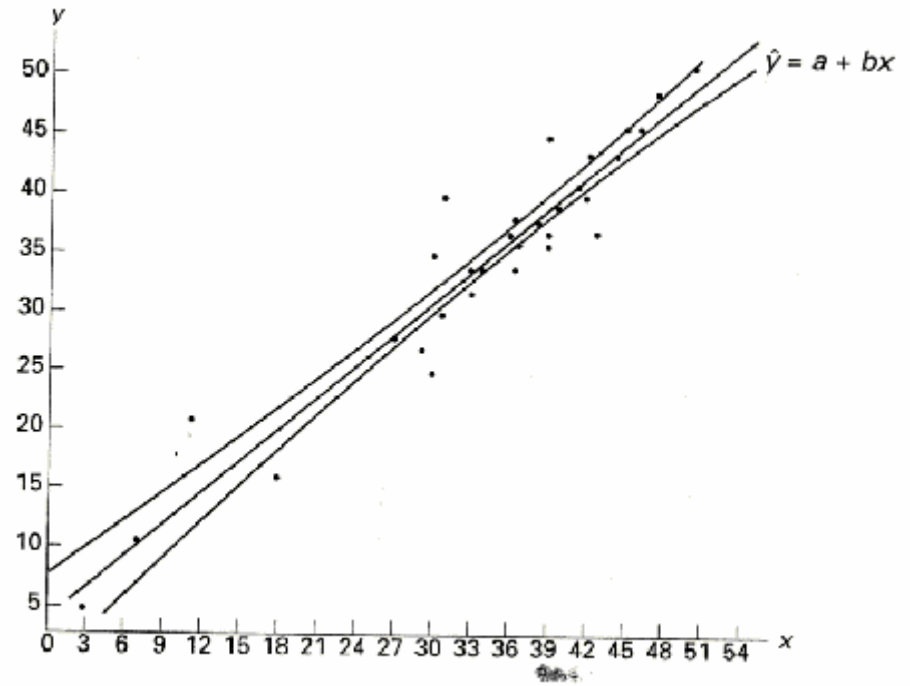


Figure 11.8 Confidence limits for the mean value of $Y|x$.

Prediction Interval for y_o, the future value of a response

A (1-α)100% prediction interval for a single response y_o in the regression line $\mu_{Y|x} = \alpha + \beta x$ is

$$\hat{y}_o - t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}} < y_o < \hat{y}_o + t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}}$$

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

Example 11.7 Using the data of Table 11.1, construct a 95% prediction interval for y_0 when $x_0 = 20\%$.

Solution We have $n = 33$, $x_0 = 20$, $\bar{x} = 33.4545$, $\hat{y}_0 = 21.9025$, $S_{xx} = 4152.18$, $s = 3.2295$, and $t_{0.025} \simeq 2.045$ for 31 degrees of freedom. Therefore, a 95% prediction interval for y_0 is

$$21.9025 - (2.045)(3.2295)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} < y_0$$

$$< 21.9025 + (2.045)(3.2295)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}},$$

which simplifies to $15.0585 < y_0 < 28.7464$.

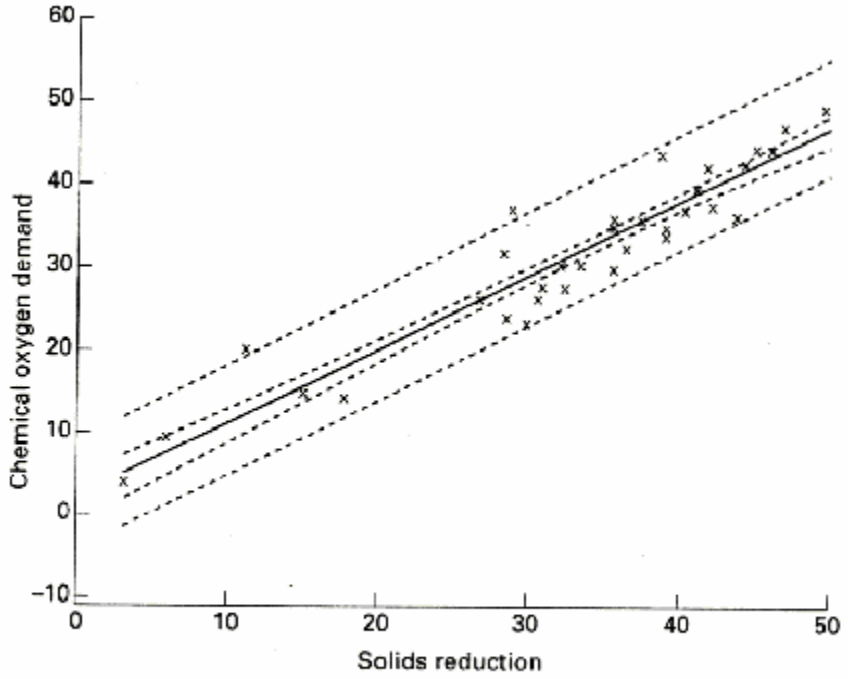


Figure 11.9 Confidence and prediction intervals for the chemical oxygen demand data.

TABLE 11.7 *Density and Stiffness for 30 Particleboards*

Density, x	Stiffness, y	Density, x	Stiffness, y
9.50	14,814.00	15.40	25,312.00
8.40	17,502.00	15.00	26,222.00
9.80	14,007.00	14.50	22,148.00
11.00	19,443.00	14.80	26,751.00
8.30	7,573.00	13.60	18,036.00
9.90	14,191.00	25.60	96,305.00
8.60	9,714.00	23.40	104,170.00
6.40	8,076.00	24.40	72,594.00
7.00	5,304.00	23.30	49,512.00
8.20	10,728.00	19.50	32,207.00
17.40	43,243.00	21.20	48,218.00
15.00	25,319.00	22.80	70,453.00
15.20	28,028.00	21.70	47,661.00
16.40	41,792.00	19.80	38,138.00
16.70	49,499.00	21.30	53,045.00

The simple linear regression fit to the data produced the fitted model

$$\hat{y} = -25{,}433.739 + 3884.976\,x \quad (R^2 = 0.7975).$$

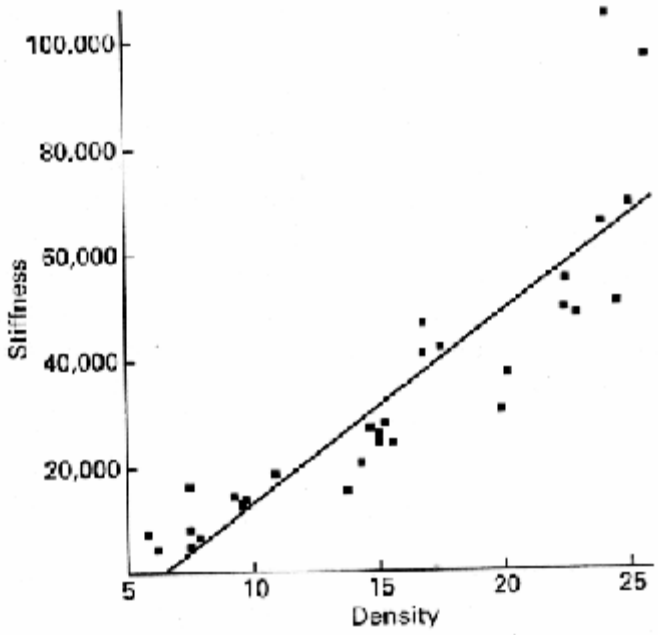


Figure 11.18 Scatterplot of the wood density data.

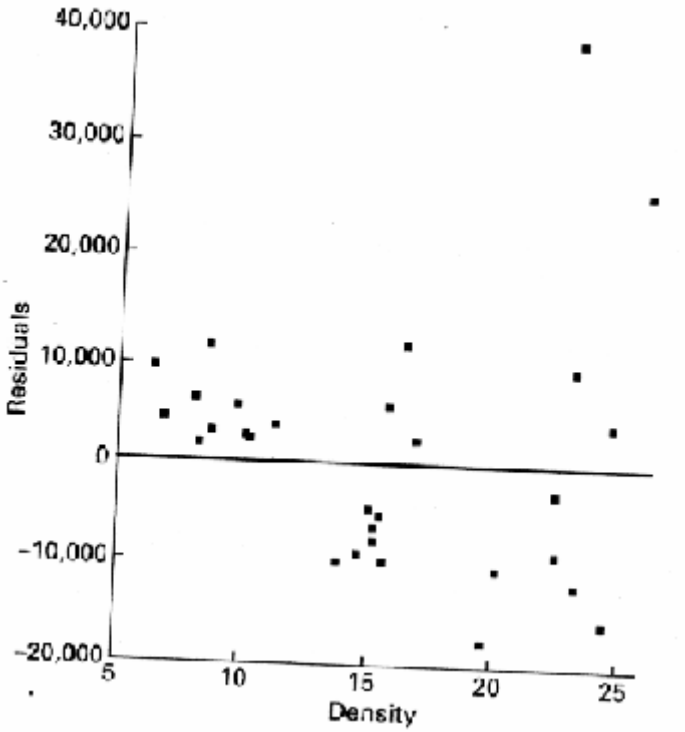


Figure 11.19 Residual plot for the wood density data.

A more complicated model may be more appropriate

$$\widehat{\ln y} = 8.257 + 0.125\,x \quad (R^2 = 0.9016).$$

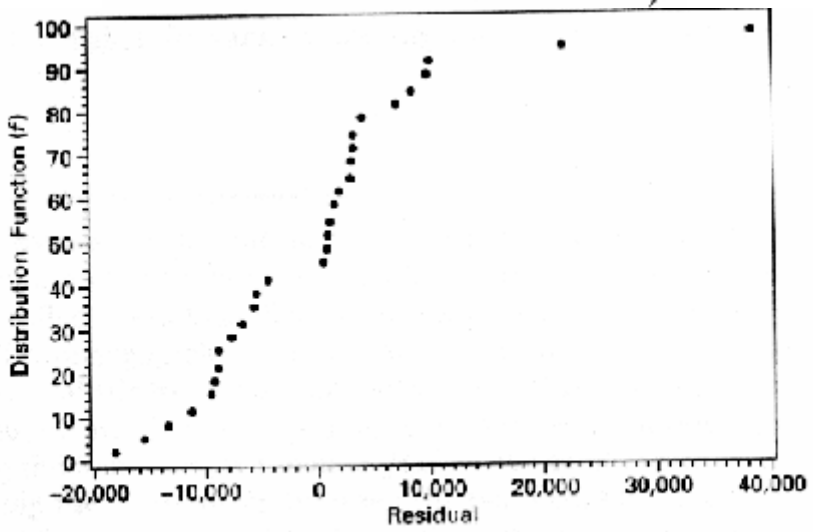


Figure 11.20 Normal probability plot of residuals for wood density data.

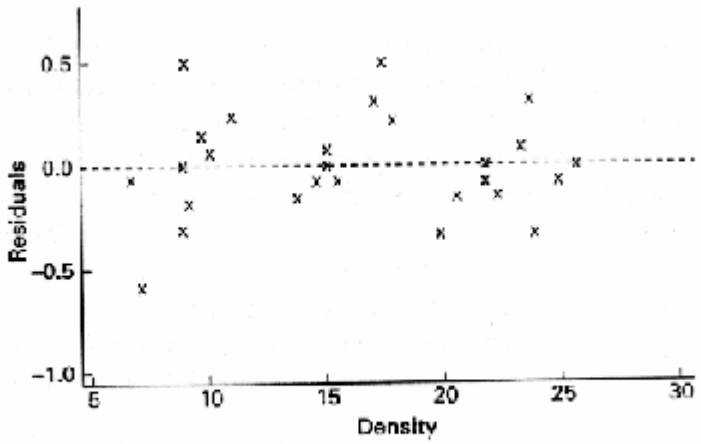


Figure 11.21 Residual plot using the log transformation for the wood density data.

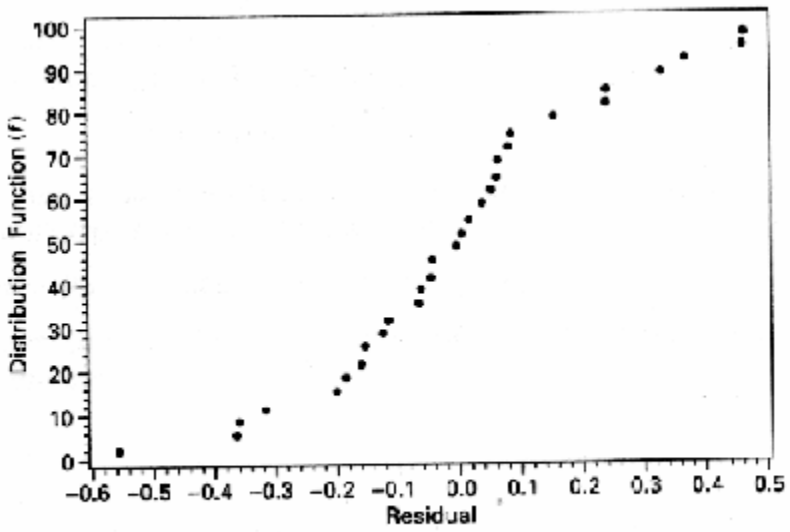


Figure 11.22 Normal probability plot of residuals for log wood density data

The higher R^2 value would suggest that the transformed model is more appropriate.

Section 11.12 Correlation

Correlation coefficient

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

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

$$-1 < r < 1$$

$r = 1.0$ or -1.0 perfect linear relationship
$r = 0.0$ no linear relationship

Sample coefficient of determination

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SSR}{S_{yy}} = \frac{SSR}{TSS}$$

represents the proportion of the variation in TSS explained by the regression of Y on x, namely, SSR.

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. According to 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 the Virginia Polytechnic Institute and State University, an experiment in which 29 loblolly pines were randomly selected for investigation yielded the data of Table 11.8 on the specific gravity in grams/cm^3 and the modulus of rupture in kilopascals (kPa). Compute and interpret the sample correlation coefficient.

TABLE 11.8

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

$$S_{xx} = 0.11273, \qquad S_{yy} = 11{,}807{,}324{,}810. \qquad S_{xy} = 34{,}422.75972.$$

Therefore,

$$r = \frac{34{,}422.275}{\sqrt{(0.11273)(11{,}807{,}324{,}810)}} = 0.9435.$$

A correlation coefficient of 0.9435 indicates a good linear relationship between X and Y. Since $r^2 = 0.8902$, we can say that approximately 89% of the variation in the values of Y is accounted for by a linear relationship with X. ⊣

Hypothesis testing on the correlation coefficient, ρ

To test the null hypothesis H_o: $\rho = \rho_o$ against any suitable alternatives,

We can use t-distribution with n-2 degrees of freedom to define the critical region, and the following test statistic

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

but this test has problem when ρ *values* close to 0 or 1

However, a more general test is given by

$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ the approximate normal distribution with mean $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ and variance $1/(n-3)$.

We can use the standard normal to define the critical region, and the following test statistic

$$z = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_o}{1-\rho_o}\right)}{1/\sqrt{n-3}} = \frac{\sqrt{n-3}}{2}\left[\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho_o}{1-\rho_o}\right)\right]$$

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

Solution

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

Example 11.12 For the data of Example 11.10 test the null hypothesis that $\rho = 0.9$ against the alternative that $\rho > 0.9$. Use a 0.05 level of significance.

Solution

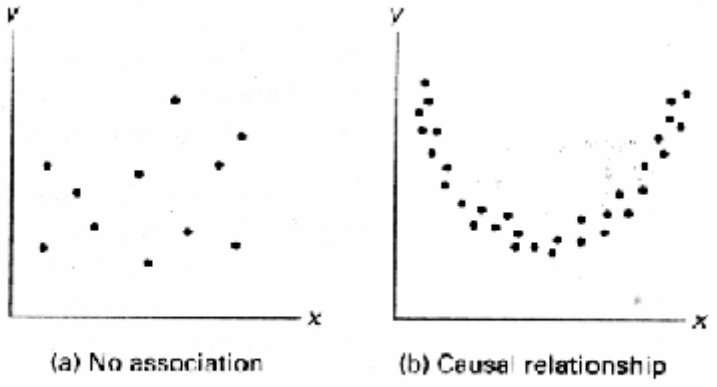
1. H_0: $\rho = 0.9$.
2. H_1: $\rho > 0.9$.
3. $\alpha = 0.05$.
4. Critical region: $z > 1.645$.
5. Computations:

$$z = \frac{\sqrt{26}}{2} \ln \left[\frac{(1 + 0.9435)0.1}{(1 - 0.9435)1.9} \right] = 1.51,$$

$$P = 0.0655$$

6. Decision: There is certainly some evidence that the correlation coefficient does not exceed 0.9. —|

Important: Correlation is a measure of linear relationship, so r = 0 does not necessarily mean there is no relationship between two variables.



(a) No association (b) Ceusal relationship

Figure 11.23 Scatter diagram showing zero correlation.