Chapter 11

Simple Linear Regression and Correlation

Chapter 11: Simple Linear Regression and Correlation

Learning objectives

1. Empirical Models
2. Simple Linear Regression
3. Properties of the Least Squares Estimators
4. Hypothesis Tests in Simple Linear Repression
5. Correlation

Empirical models

- Many problems in engineering and science involve exploring the relationships between two or more variables.
- **Regression analysis** is a statistical technique that is very useful for these types of problems.
- For example, in a chemical process, suppose that the yield of the product is related to the process-operating temperature.
- Regression analysis can be used to build a model to predict yield at a given temperature level.

Empirical models

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Empirical models

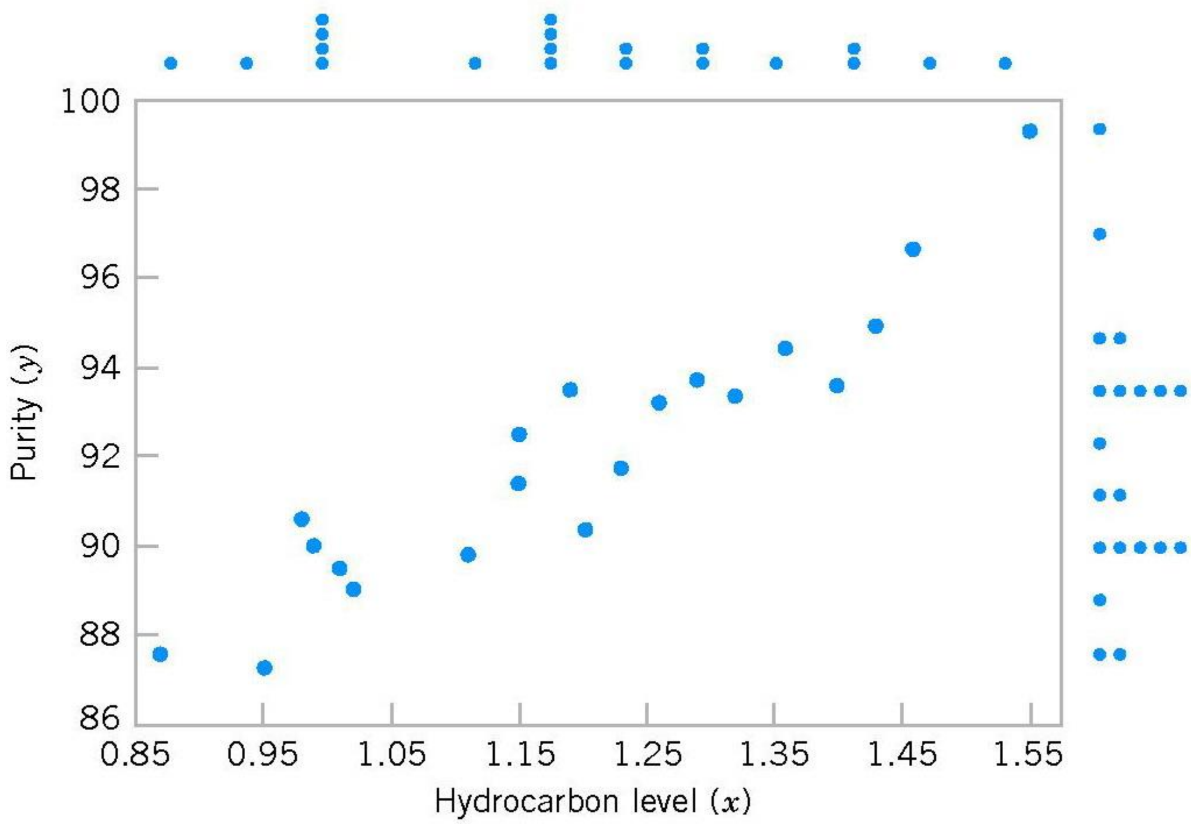


Figure 11-1 Scatter Diagram of oxygen purity versus hydrocarbon level from Table 11-1.

Empirical models

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable *Y* is related to *x* by the following straight-line relationship:

$$E(Y \mid x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

regression coefficients.

The **simple linear regression model** is given by

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

random error

Empirical models

Suppose that the mean and variance of ε are 0 and σ^2, respectively, then

$$E(Y|x) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$$

The variance of *Y* given *x* is

$$V(Y|x) = V(\beta_0 + \beta_1 x + \varepsilon) = V(\beta_0 + \beta_1 x) + V(\varepsilon) = 0 + \sigma^2 = \sigma^2$$

The true regression model is a line of mean values:

$$\mu_{Y|x} = \beta_0 + \beta_1 x$$

Simple Linear Regression

- The case of simple linear regression considers a single regressor or predictor x and a dependent or response variable Y.

- The expected value of Y at each level of x is a random variable:

$$E(Y \mid x) = \beta_0 + \beta_1 x$$

- We assume that each observation, Y, can be described by the model

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Simple Linear Regression

Suppose that we have n pairs of observations (x_1, y_1), (x_2, y_2), …, (x_n, y_n):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \; i = 1, \ldots, n$$

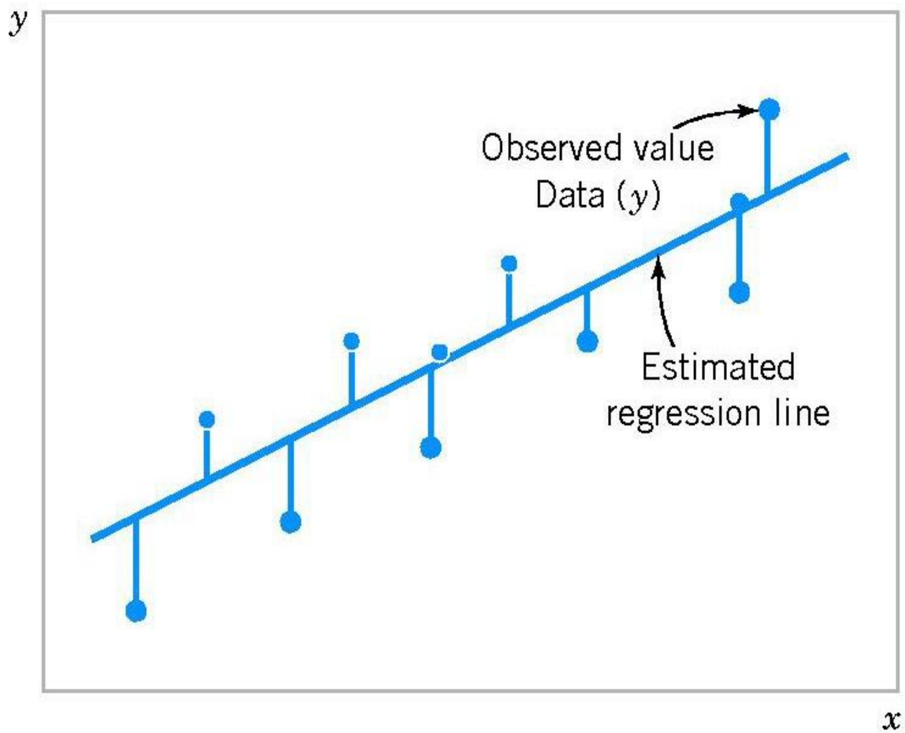


Figure 11-3 Deviations of the data from the estimated regression model.

Simple Linear Regression

The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

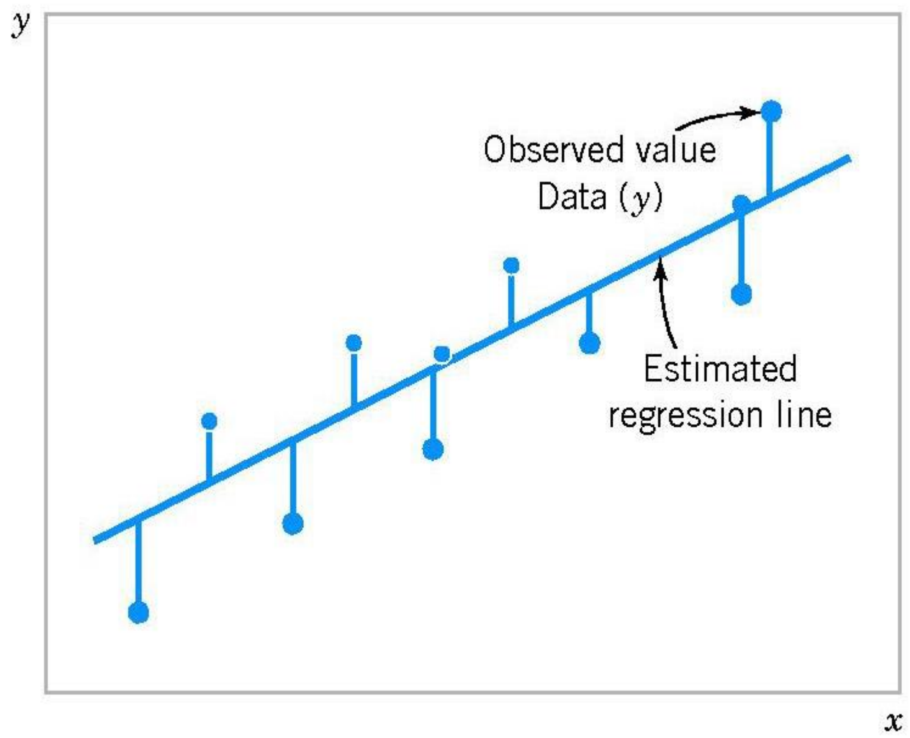


Figure 11-3
Deviations of the data from the estimated regression model.

Simple Linear Regression

Simplifying these two equations yields

$$\left.\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned}\right\} \boxed{\hat{\beta}_0,\ \hat{\beta}_1 = ?}$$

Notation

$$S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Simple Linear Regression

Theorem

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Estimated regression line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Simple Linear Regression

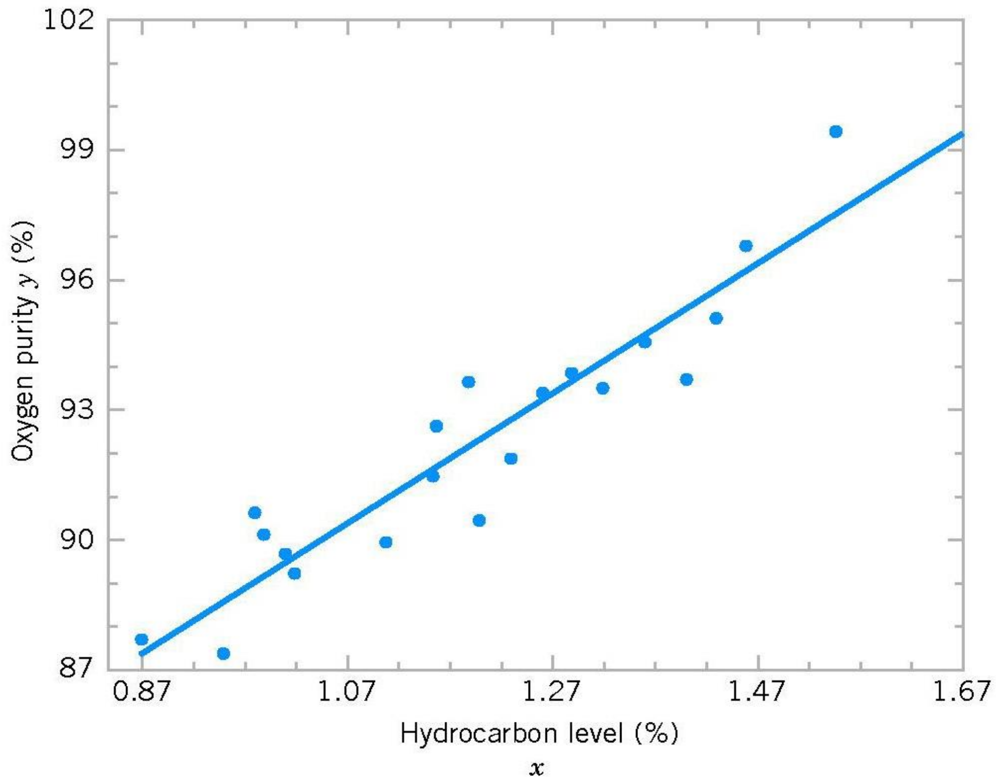
Example

We will fit a simple linear regression model to the oxygen purity data in Table 11-1. The following quantities may be computed:

Simple Linear Regression

The fitted simple linear regression model is

$$\hat{y} = 74.283 + 14.947x$$



Simple Linear Regression

Estimating σ^2

The error sum of squares is

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

We have

$$E(SS_E) = (n - 2)\sigma^2.$$

$$SS_E = SS_T - \hat{\beta}_1 S_{xy}$$

Where $SS_T = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

Estimating σ²

Estimating σ²

Theorem

An **unbiased estimator** of σ² is

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

Standard error

where

$$SS_E = SS_T - \hat{\beta}_1 S_{xy}$$

11.3. Properties of the Least Squares Estimator

11.3. Properties of the Least Squares Estimator

Mean and Variance of Estimators

$$E(\hat{\beta}_1) = \beta_1$$

$$E(\hat{\beta}_0) = \beta_0$$

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}$$

Hypothesis Tests in Simple Linear Regression

Test on the β_1

$$H_0: \beta_1 = \beta_{1,0}$$
$$H_1: \beta_1 \neq \beta_{1,0}$$

Test statistic

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

has the t distribution with n - 2 degrees of freedom.

If $|t_0| > t_{\alpha/2,\, n-2}$: reject H_0

If $|t_0| < t_{\alpha/2,\, n-2}$: fail to reject H_0

Hypothesis Tests in Simple Linear Regression

Test on the β_1

An important special case

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

These hypotheses relate to the **significance of regression**.

Failure to reject H_0 is equivalent to concluding that there is no linear relationship between x and Y.

Hypothesis Tests in Simple Linear Regression

Test on the β_1

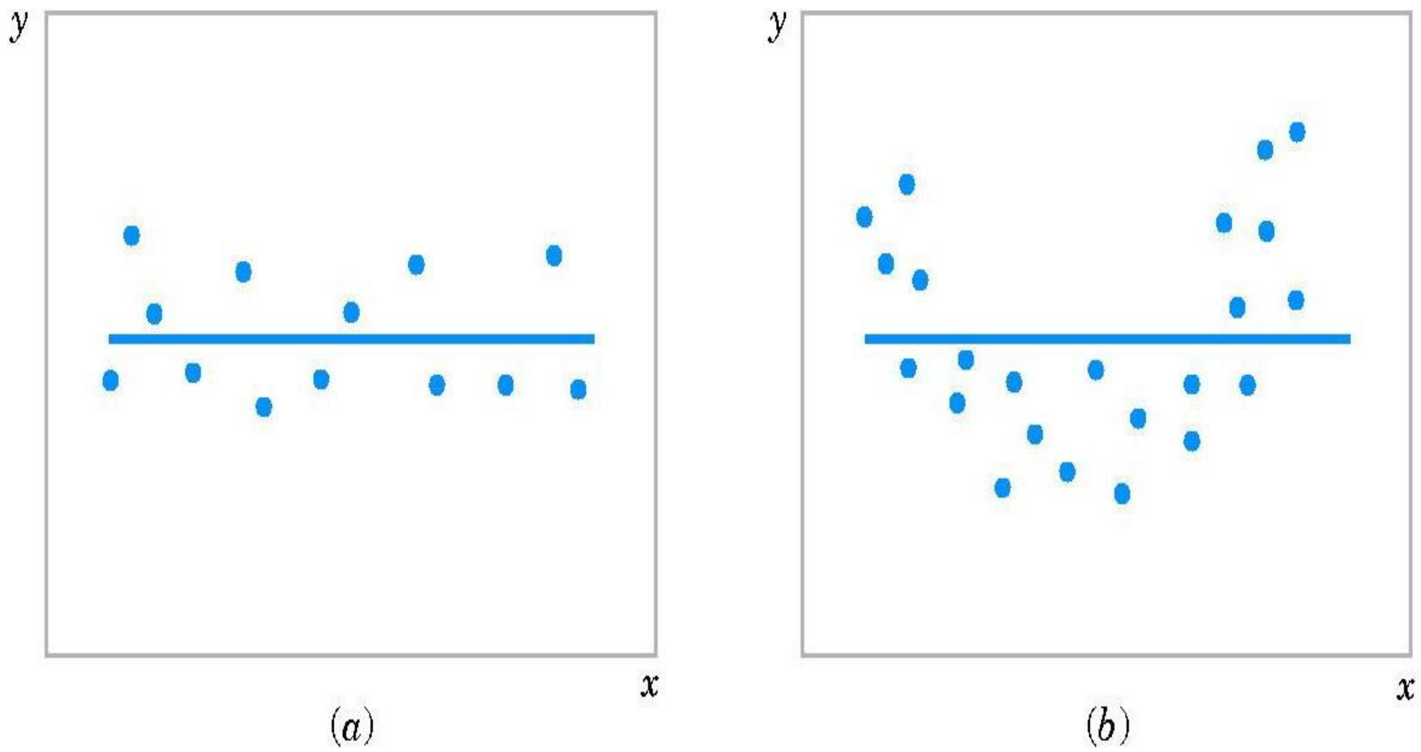


Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

Hypothesis Tests in Simple Linear Regression

Test on the β_1

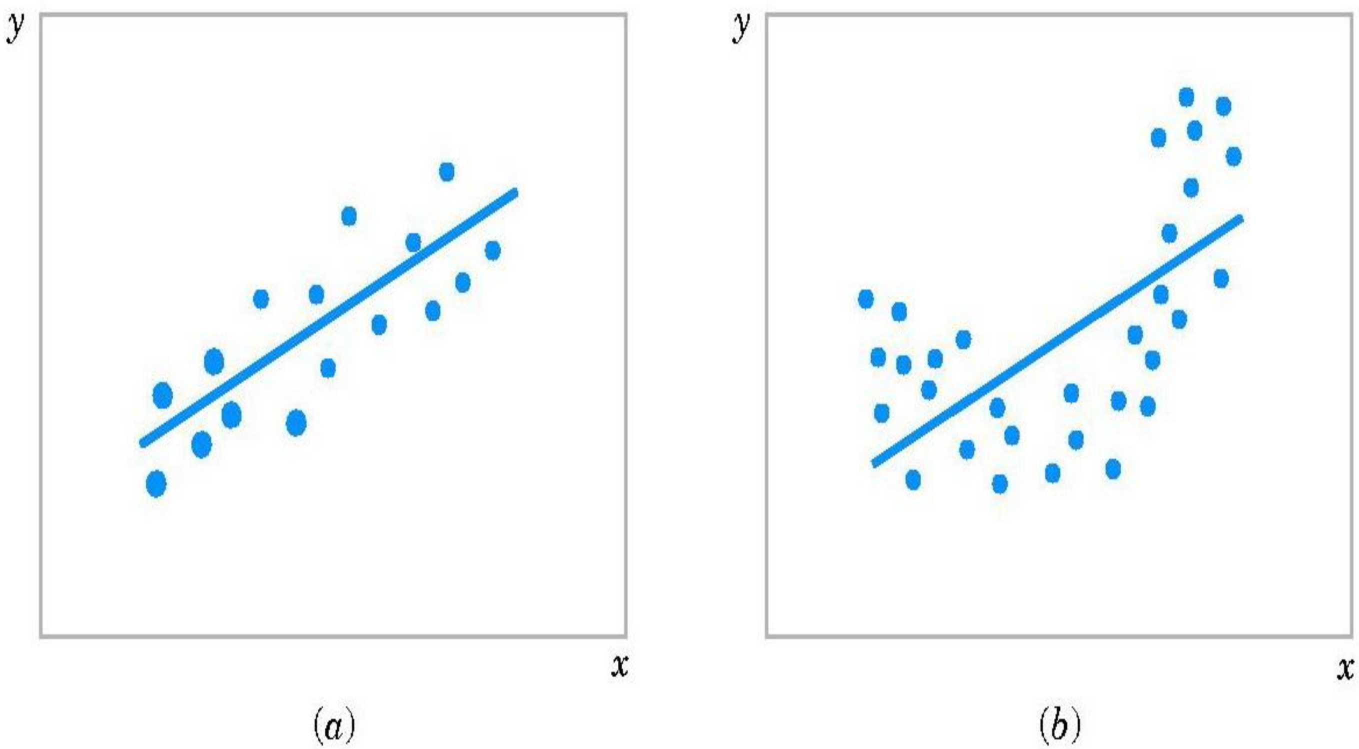


Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

Hypothesis Tests in Simple Linear Regression

Example

We will test for significance of regression using the model for the oxygen purity data from Table 11-1. The hypotheses are

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

and we will use $\alpha = 0.01$.

Hypothesis Tests in Simple Linear Regression

Test on the β_0

$$H_0: \beta_0 = \beta_{0,0}$$
$$H_1: \beta_0 \neq \beta_{0,0}$$

Test statistic

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

If $|t_0| > t_{\alpha/2,\, n-2}$: reject H_0

If $|t_0| < t_{\alpha/2,\, n-2}$: fail to reject H_0

Confidence Intervals

Confidence Intervals on the Slope and Intercept

Under the assumption that the observations are normally and independently distributed, a 100(1-α)% confidence interval on the slope β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Similarly, a 100(1-α)% confidence interval on the intercept β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

Confidence Intervals

Example

We will find a 95% confidence interval on the slope of the regression line using the data in Table 11-1.

Confidence Intervals

Confidence Interval on the Mean Response

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

A 100(1-α)% confidence interval about the mean response at the value of $x=x_0$ is given by

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

$$\le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

Confidence Intervals

Example

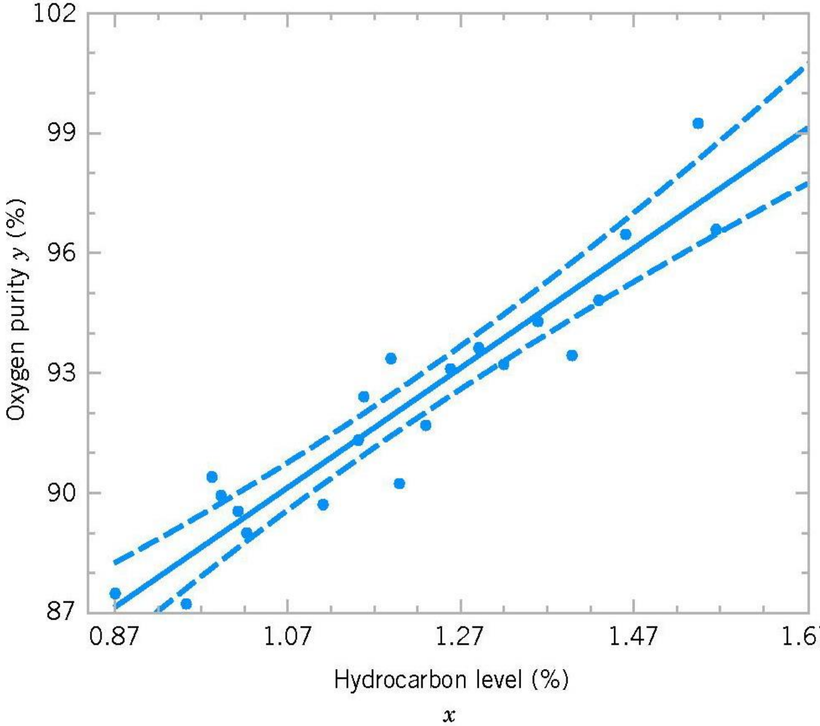
We will find a 95% confidence interval about the mean response for the data in Table 11-1.

Confidence Intervals

$$\left\{89.23 \pm 2.101\sqrt{1.18\left[\frac{1}{20}+\frac{(1.00-1.1960)^2}{0.68088}\right]}\right\}$$

$$88.48 \le \mu_{Y|1.00} \le 89.98$$

Scatter diagram of oxygen purity with fitted regression line and 95% confidence limits on $\mu_{Y|x0}$.



Prediction of New Observations

A 100(1-α)% prediction interval on a future observation Y_0 at the value x_0 is given by

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]}$$

Correlation

Definition

The **sample correlation coefficient**

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}\, SS_T}}$$

Note that

$$\hat{\beta}_1 = \left(\frac{SS_T}{S_{XX}}\right)^{1/2} R$$

We may also write:

$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} = \frac{\hat{\beta}_1 S_{XY}}{SS_T} = \frac{SS_R}{SS_T}$$

Correlation

Properties of the Linear Correlation Coefficient *r*

1. $-1 \le r \le 1$
2. The value of r does not change if all values of either variable are converted to a different scale.
3. The value of r is not affected by the choice of x and y. Interchange all x- and y-values and the value of r will not change.
4. r measures strength of a linear relationship.

Correlation

Example 11.10:

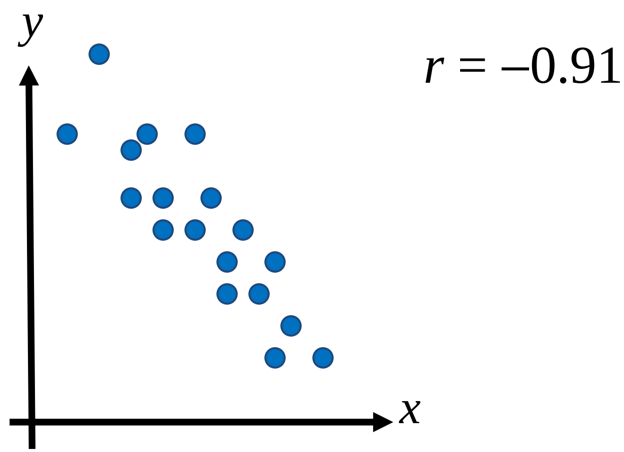
It is important that scientific researchers in the area of forest products be able to study correlation among the anatomy and mechanical properties of trees. For the study *Quantitative Anatomical Characteristics of Plantation Grown Loblolly Pine (Pinus Taeda L.) and Cottonwood (Populus deltoides Bart. 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/cm3 and the modulus of rupture in kilopascals (kPa). Compute and interpret the sample correlation coefficient.

Example 11.10:

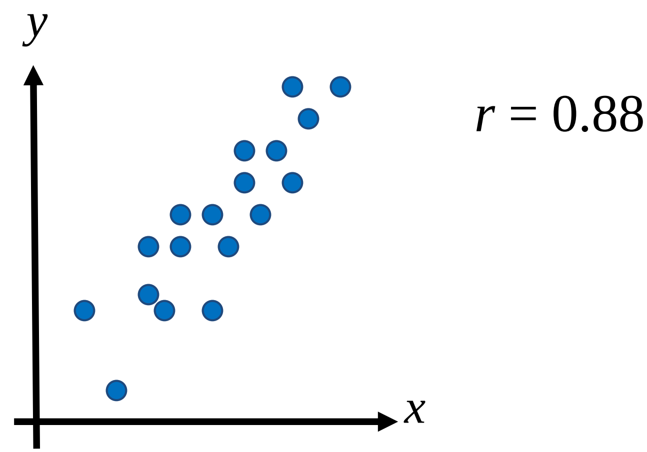
Table 11.9: Data on 29 Loblolly Pines for Example 11.10

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		

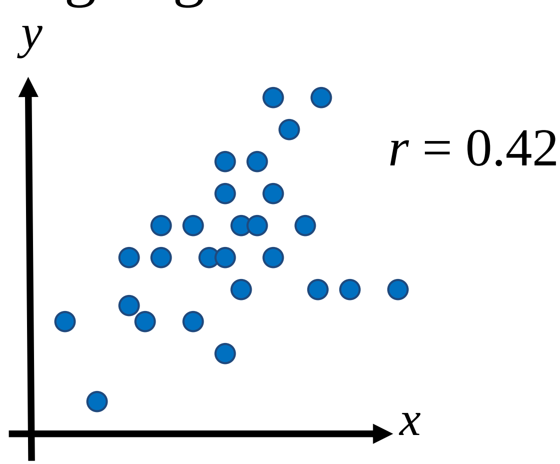
Correlation



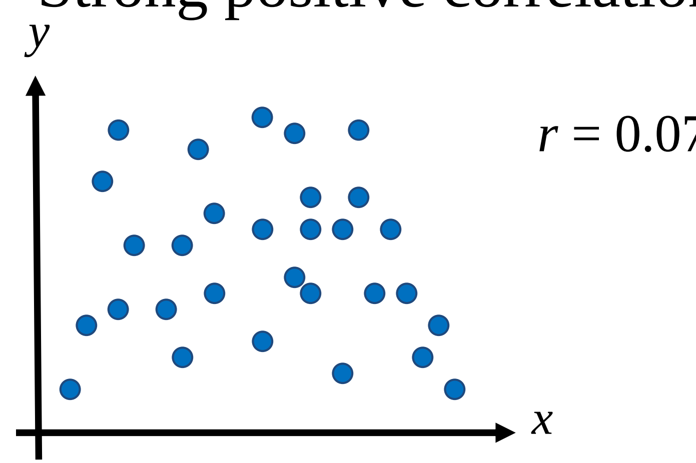
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

Correlation

Test on the ρ

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

Test statistic $$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

has the *t* distribution with *n* - 2 degrees of freedom.

If $|t_0| > t_{\alpha/2,\, n-2}$: reject H_0

If $|t_0| < t_{\alpha/2,\, n-2}$: fail to reject H_0

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

Correlation

Test on the ρ

$H_0: \rho = \rho_0$

$H_1: \rho \neq \rho_0$

Test statistic $Z_0 = (\operatorname{arctanh} R - \operatorname{arctanh} \rho_0)\sqrt{n-3}$

$\boxed{\tanh u = (e^u - e^{-u})/(e^u + e^{-u})}$

If $|t_0| > z_{\alpha/2}$: reject H_0

If $|t_0| < z_{\alpha/2}$: fail to reject H_0