Chapter 11: Simple Linear Regression and Correlation

Learning objectives

1. Empirical Models
2. Simple Linear Regression
3. 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

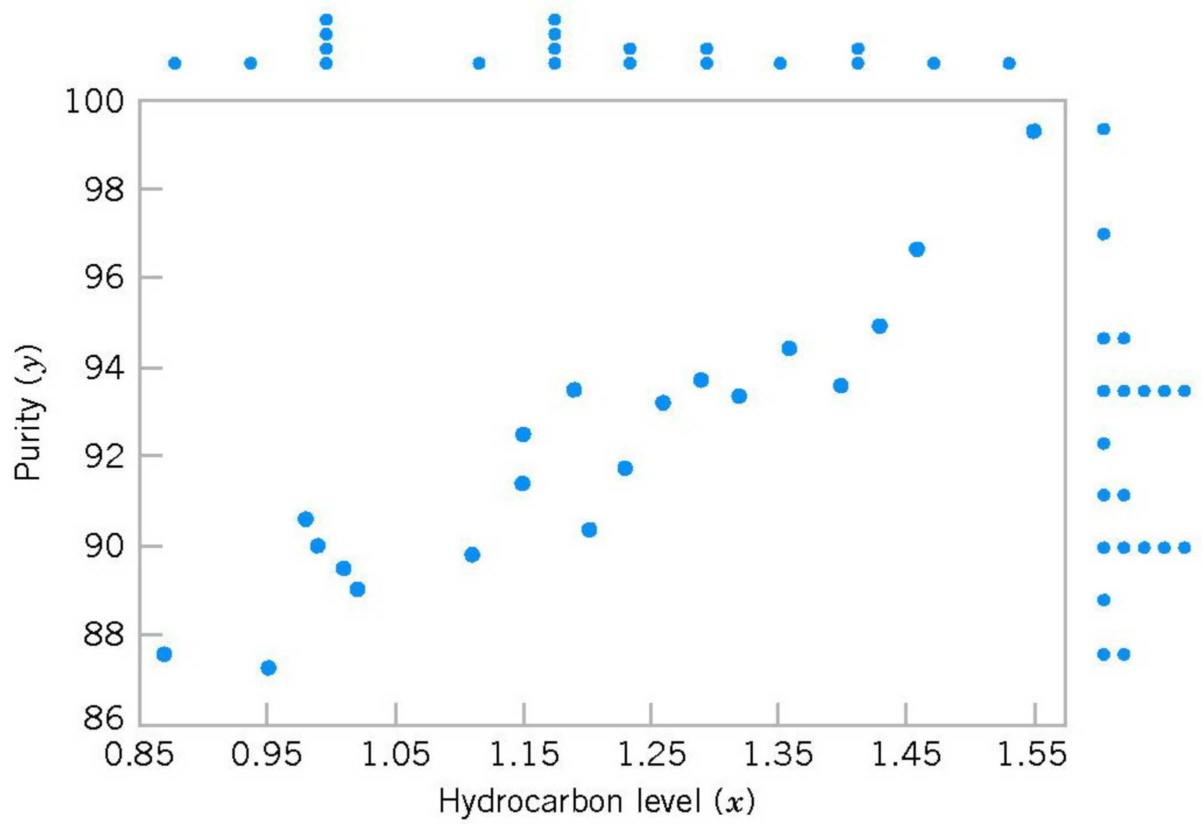


Figure 11-1 Scatter Diagram of oxygen purity versus hydrocarbon level from Table 11-1.

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:

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

Empirical Models

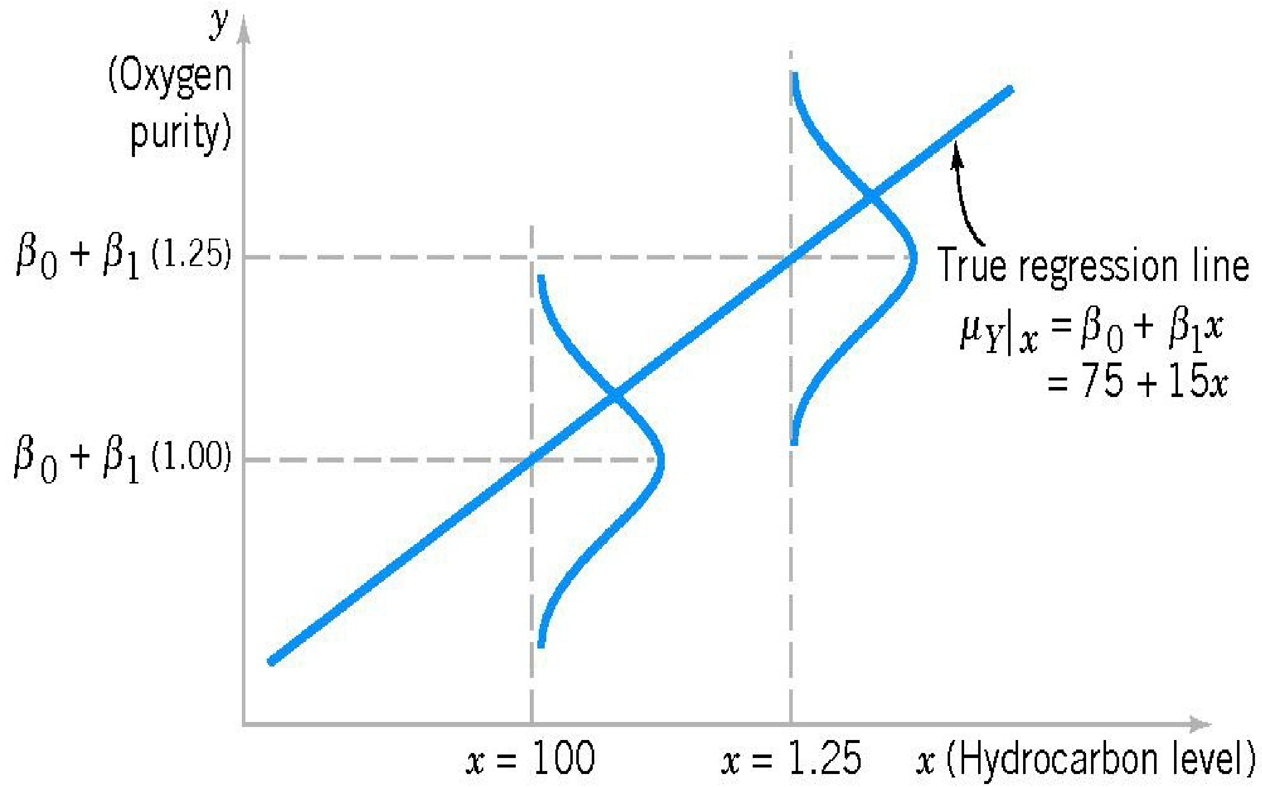


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity-hydrocarbon data.

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):

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

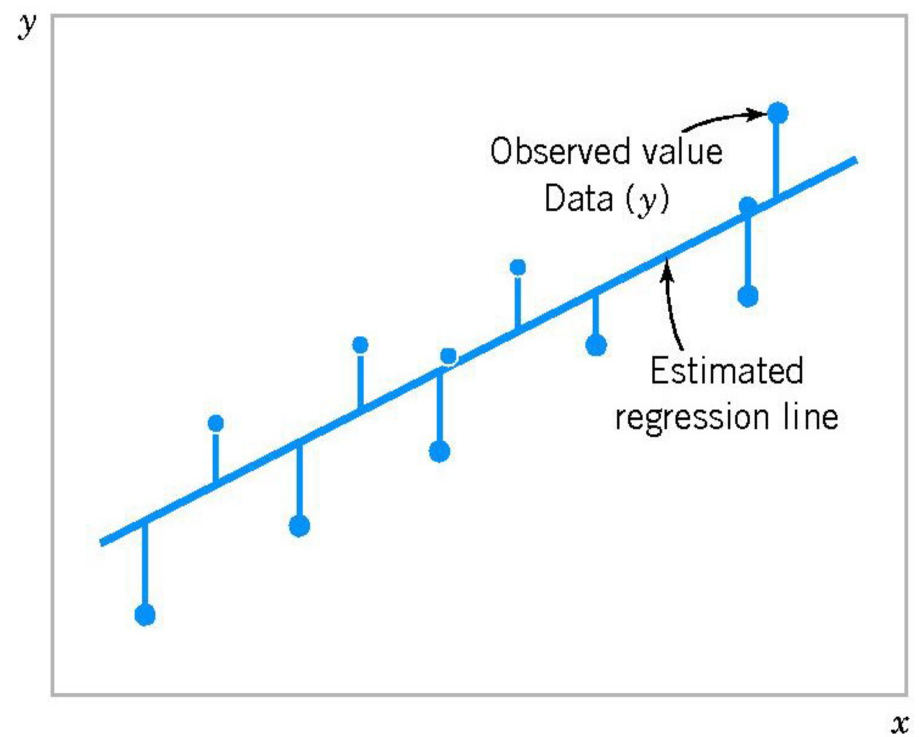
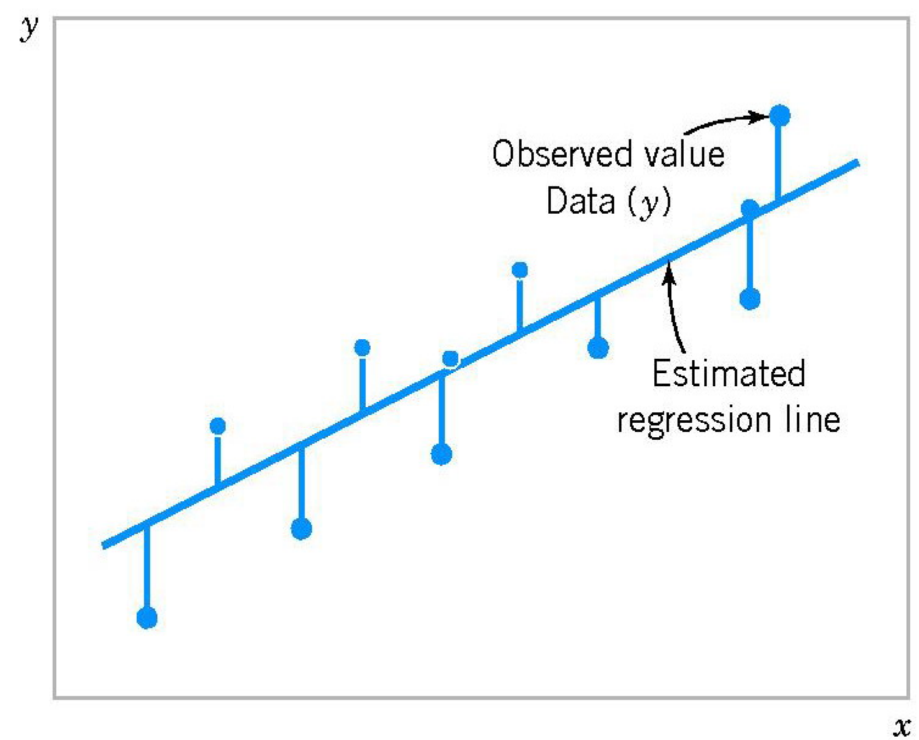


Figure 11-3
Deviations of the data from the estimated regression model.

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

The sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$$

The least squares estimators of β_0 and β_1, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

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

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

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

Theorem

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

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

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

Estimated regression line is

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

Simple Linear Regression

Empirical models

Simple Linear Regression

Estimating σ^2

Hypothesis tests

Confidence intervals

Prediction

Adequacy

Correlation

Summary

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:

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1{,}843.21 \quad \bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170{,}044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892 \quad \sum_{i=1}^{20} x_i y_i = 2{,}214.6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20} = 0.68088$$

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} = 2{,}214.6566 - \frac{(23.92)(1{,}843.21)}{20} = 10.17744$$

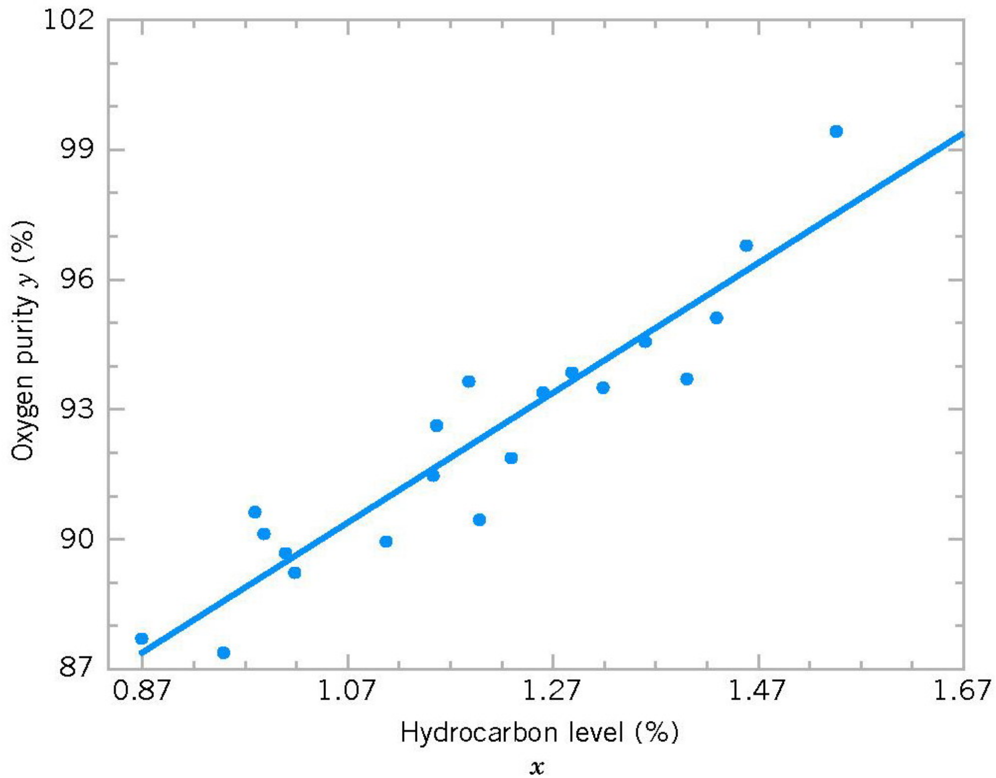
Therefore, the least squares estimates of the slope and intercept are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

The fitted simple linear regression model is

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



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

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

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

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

Estimating σ^2

Theorem

An **unbiased estimator** of σ^2 is

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

$\boxed{\hat{\sigma}}$ Standard error

where

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

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

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

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.

Test on the β_1

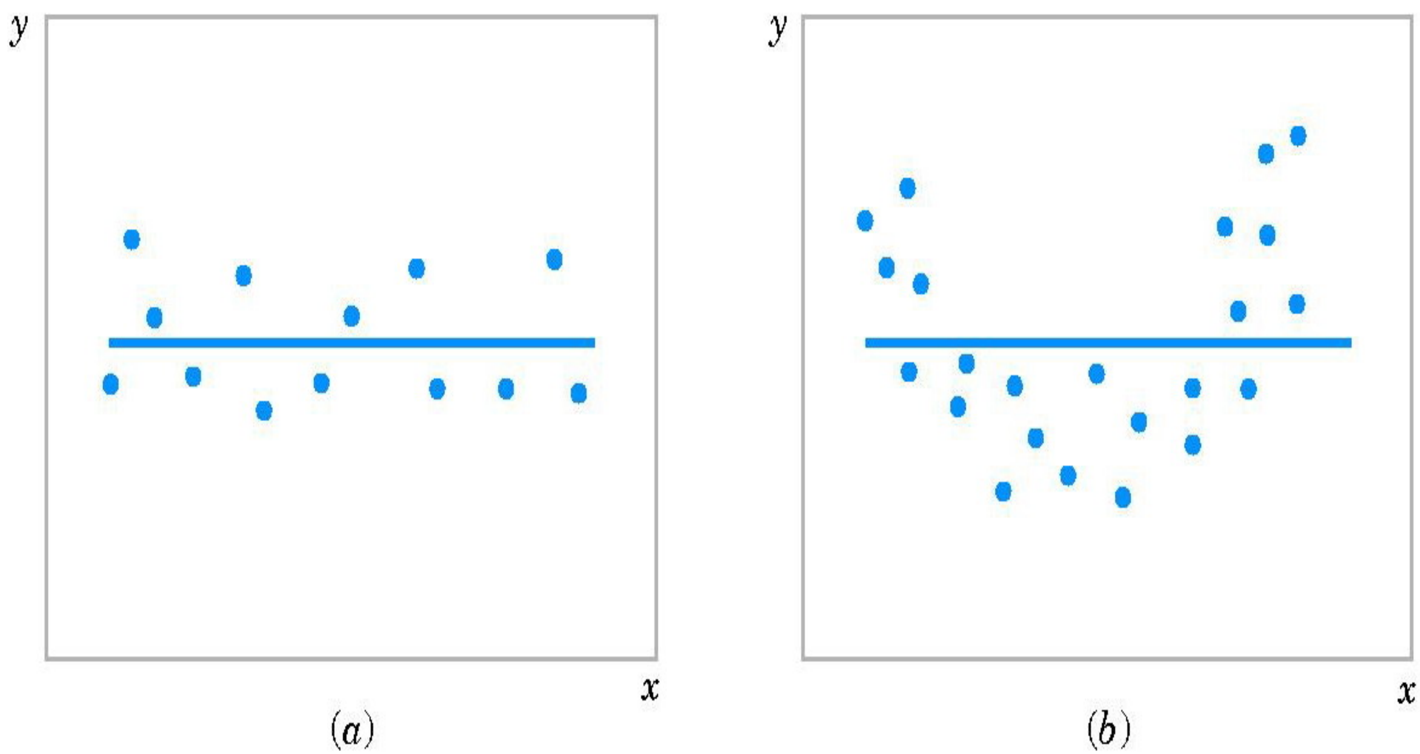


Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

Test on the β_1

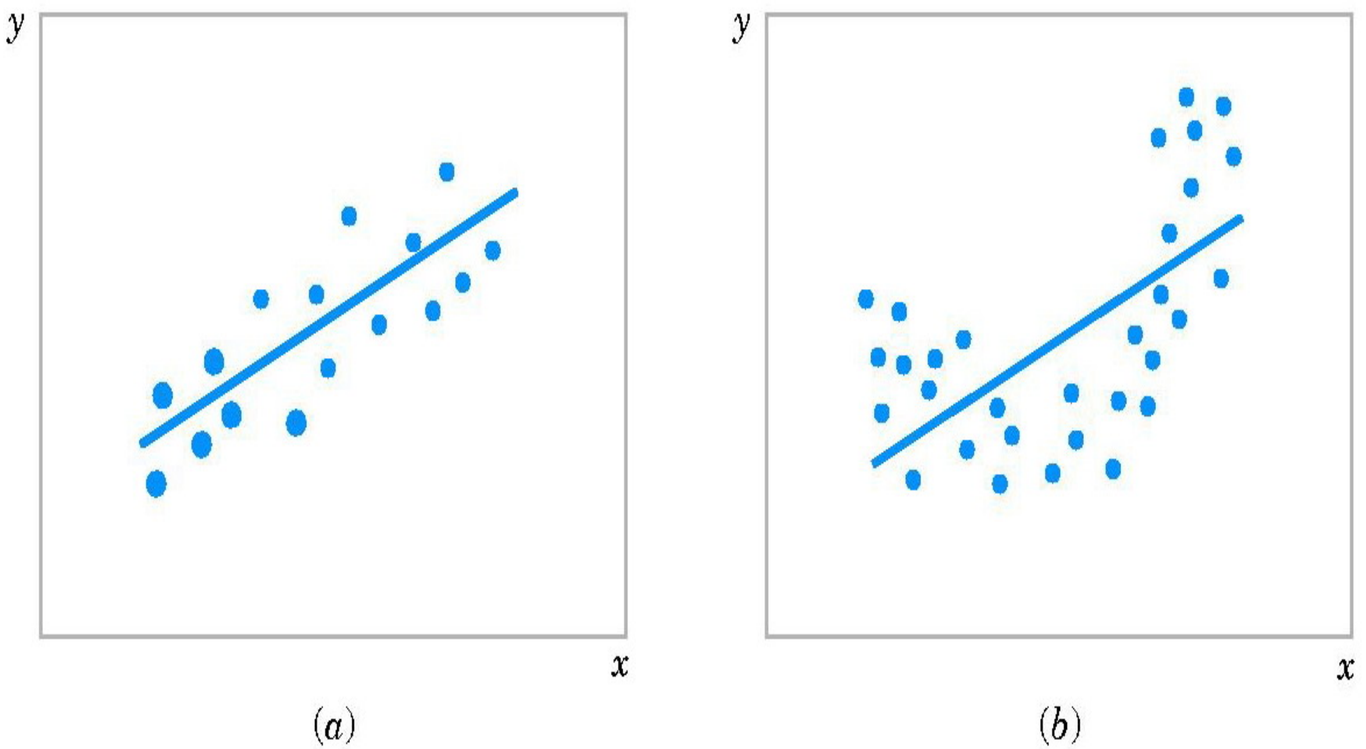


Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

Example

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

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

and we will use $\alpha = 0.01$.

Recall $\hat{\beta}_1 = 14.97 \quad n = 20, \quad S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18$

Test statistic $t_0 = \dfrac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \dfrac{14.947}{\sqrt{1.18/0.68088}} = 11.35$

Hypothesis Tests in Simple Linear Regression

$$t_{\alpha/2,\, n-2} = t_{0.005,18} = 2.88 < |t_0|$$

$$\Downarrow$$

Reject H_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

Test on the β_0

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

Test statistic

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

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

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

Confidence Intervals

Example

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

Recall $\hat{\beta}_1 = 14.947,\ S_{xx} = 0.68088,$ and $\hat{\sigma}^2 = 1.18$

CI 95% for β_1 :

$$\hat{\beta}_1 - t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$14.947 - 2.101\sqrt{\frac{1.18}{0.68088}} \le \beta_1 \le 14.947 + 2.101\sqrt{\frac{1.18}{0.68088}}$$

$$12.197 \le \beta_1 \le 17.697$$

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

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

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

Confidence Intervals

Example

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

The fitted model is $\hat{\mu}_{Y|x_0} = 74.283 + 14.947x_0,$

If we are interested in predicting mean oxygen purity when $x_0 = 100\%$ then

$$\hat{\mu}_{Y|x_{1.00}} = 74.283 + 14.947(1.00) = 89.23$$

CI 95% on $\mu_{Y|x_0}$

$$\hat{\mu}_{Y|x_0} \pm 2.101\sqrt{1.18\left[\frac{1}{20} + \frac{(x_0 - 1.1960)^2}{0.68088}\right]}$$

Confidence Intervals

Empirical models

Simple Linear Regression

Estimating σ^2

Hypothesis tests

Confidence intervals

Prediction

Adequacy

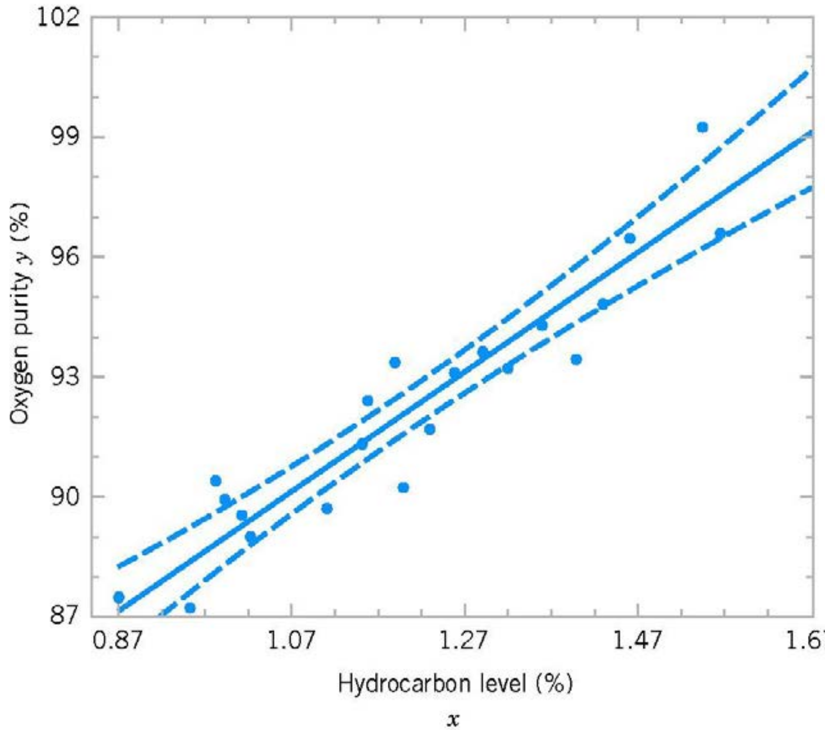
Correlation

Summary

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

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

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

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

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

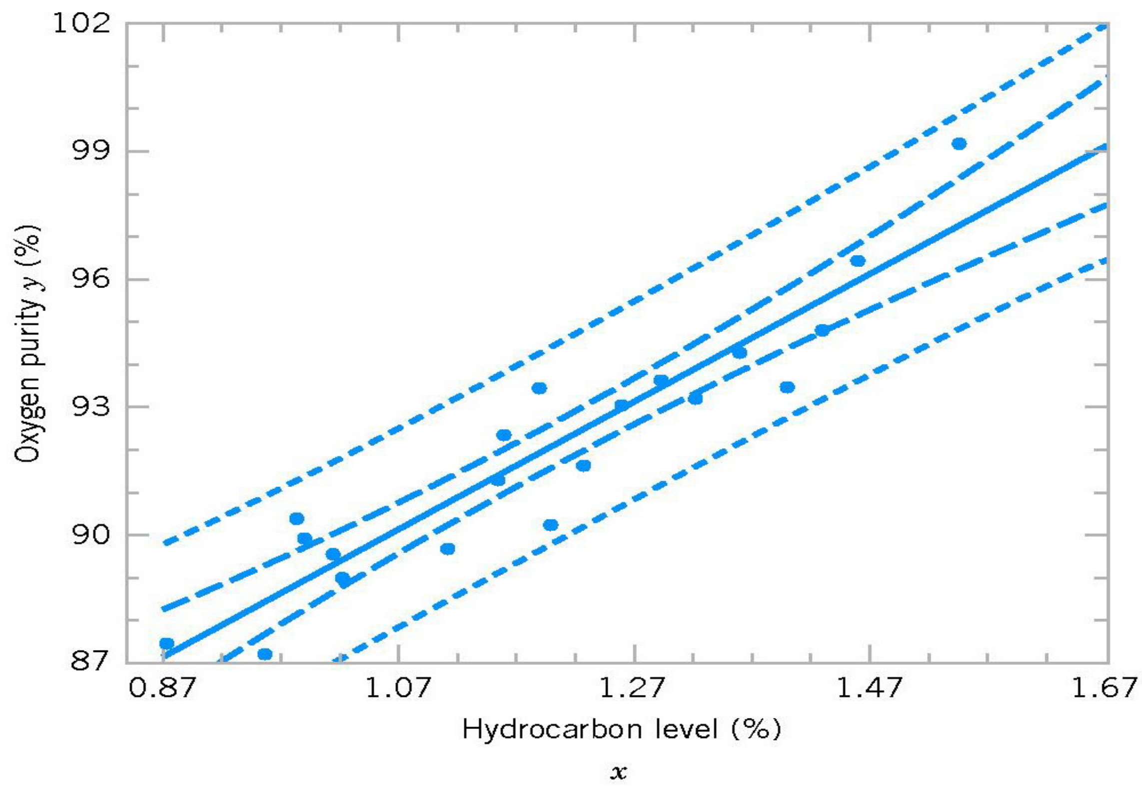
Return to Table 11.1, confidence interval 95% on Y_0 at $x_0 = 100\%$

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

$$\leq Y_0 \leq 89.23 + 2.101\sqrt{1.18\left[1+\frac{1}{20}+\frac{(1.00-1.1960)^2}{0.68088}\right]}$$

$$\boxed{86.83 \leq y_0 \leq 91.63}$$

Prediction of New Observations



Scatter diagram of oxygen purity data from Table 11.1 with fitted regression line, 95% prediction limits, and 95% confidence limits on $\mu_{Y|x0}$.

Adequacy of the Regression Model

- Fitting a regression model requires several **assumptions.**

  1. Errors are uncorrelated random variables with mean zero;
  2. Errors have constant variance; and,
  3. Errors be normally distributed.

- The analyst should always consider the validity of these assumptions to be doubtful and conduct analyses to examine the adequacy of the model

Adequacy of the Regression Model

Coefficient of Determination (R^2)

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}.$$

- $0 \le R^2 \le 1$;
- We often refer to R^2 as the amount of variability in the data explained or accounted for by the regression model.

Adequacy of the Regression Model

Example

For the oxygen purity regression model,

$$R^2 = SS_R/SS_T$$
$$= 152.13/173.38$$
$$= 0.877$$

Thus, the model accounts for 87.7% of the variability in the data.

Definition

The **sample correlation coefficient**

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

Note that

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

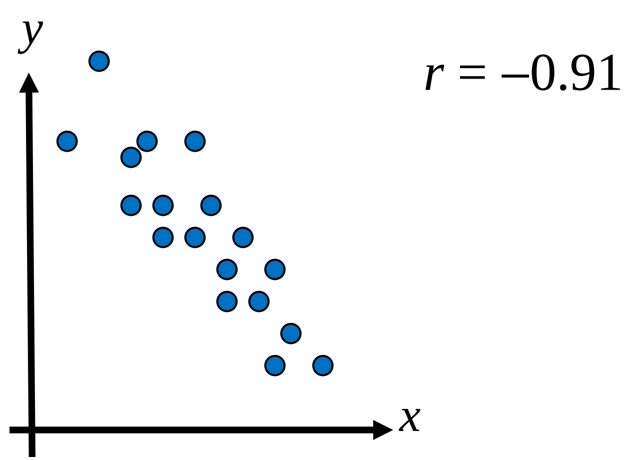
We may also write:

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

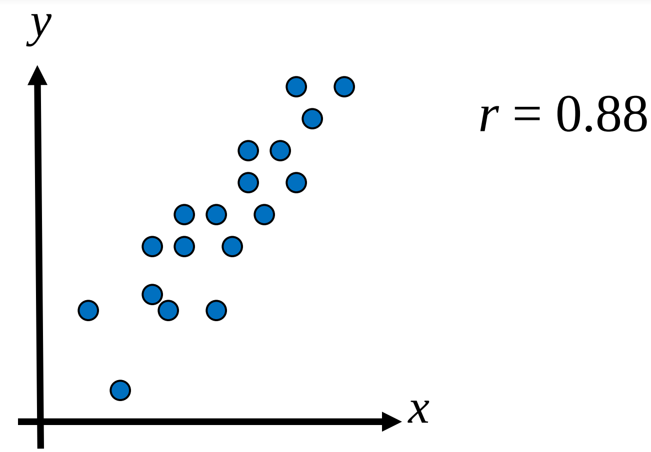
Properties of the Linear Correlation Coefficient r

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.

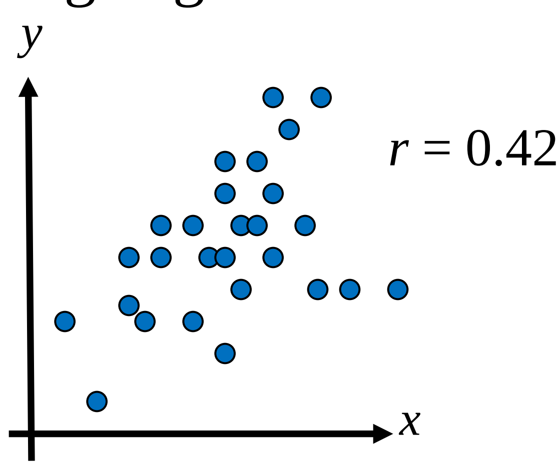
Correlation



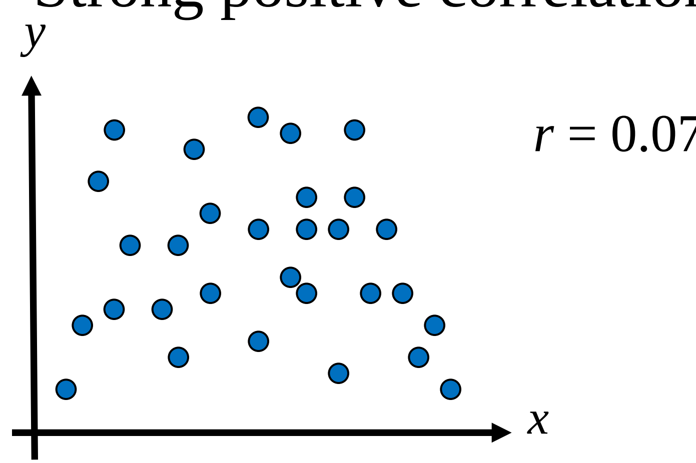
Strong negative correlation



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

Test on the ρ

$$H_0: \rho = \rho_0$$
$$H_1: \rho \neq \rho_0$$

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

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

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

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

SUMMARY

We have studied:

1. Empirical Models
2. Simple Linear Regression
3. Correlation

Homework: Read slides of the next lecture.