Simple Linear Regression & Correlation

Tim Pengajar IF2220 Probabilitas dan Statistika
Reference: Walpole Chapter 11

Introduction

Simple Linear Regression Model

Least Square Estimator

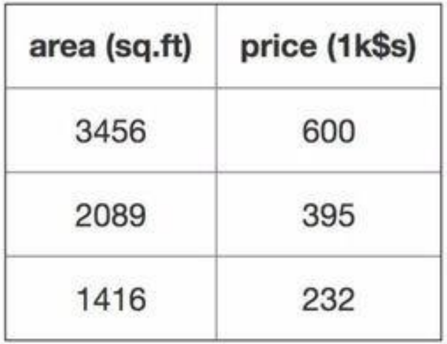
A Measure of Quality of Fit

Transformations

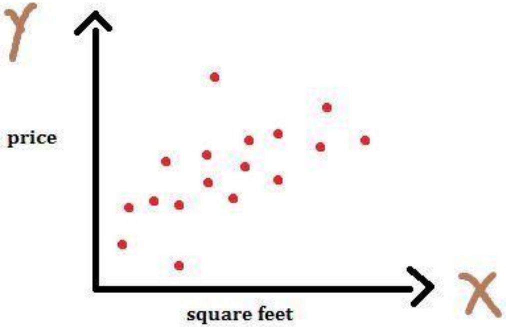
Multiple Linear Regression

Correlation

Housing price and its area



area (sq.ft)	price (1k$s)
3456	600
2089	395
1416	232



Source: https://medium.com/datadriveninvestor/understanding-linear-regression-in-machine-learning-643f577eba84

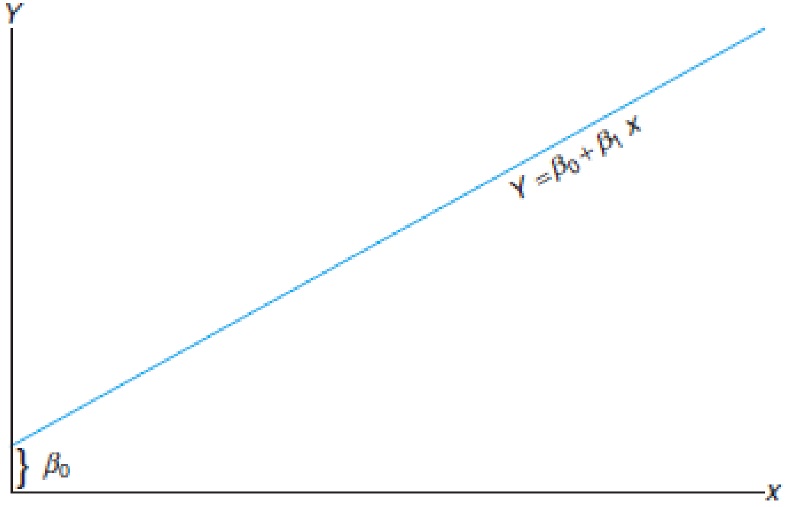
Linear Relation

Houses in the same part of the country that have the same square footage of living space will not all be sold for the same price.

The price of houses (in thousands of dollars) is natural dependent variable, or response.

Square feet of living space is natural independent variable, or regressor.

In this example, the relationship is not deterministic (i.e., a given x does not always give the same value for Y).



A reasonable form of a relationship between the response Y and the regressor x is the linear relationship: Y = β0 +β1x

β0 is the intercept and β1 is the slope

Regression Analysis

The concept of regression analysis deals with finding the best relationship between Y and x, quantifying the strength of that relationship, and using methods that allow for prediction of the response values given values of the regressor x.

Simple linear regression treats only the case of a single regressor variable in which the relationship between y and x is linear.

Relationships among variables are not deterministic (i.e., not exact). There must be a random component to the equation that relates the variables.

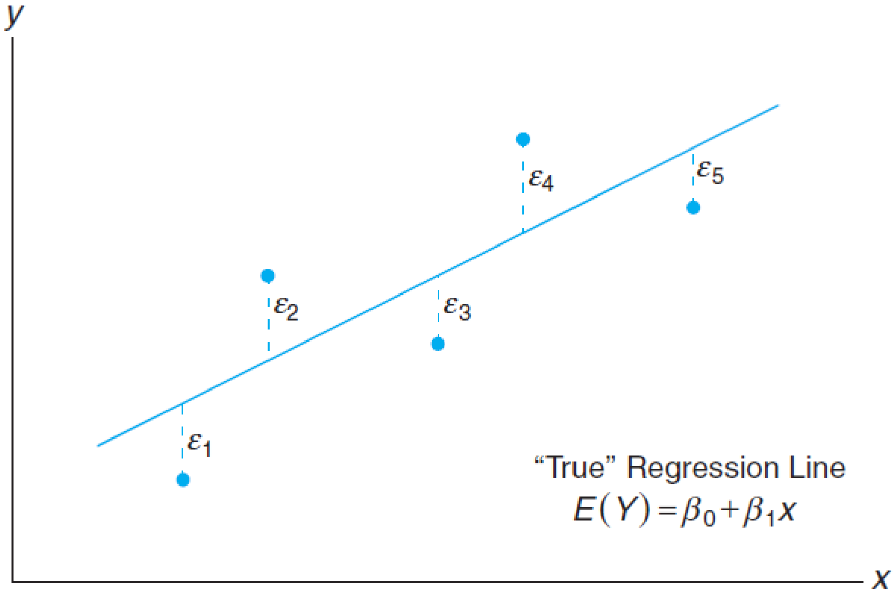
Simple Linear Regression Model

Y = β0 + β1x + **ε**

β0: unknown intercept,
β1: unknown slope,
ε: a random variable that is assumed to be distributed with E(**ε**) = 0 and Var(**ε**) = σ^2.

Y is a random variable since **ε** is random.
Eg. Y = house price

X is not random, and is measured with negligible error, eg. X = living space



Hypothetical (x, y) data scattered around the true regression line for n = 5. We never observe the actual values in practice and thus we can never draw the true regression line (but we assume it is there).

Fitted Regression Line

An important aspect of regression analysis is to estimate the parameters $\beta0$ and $\beta1$ (i.e., estimate regression coefficients/parameters).

Suppose we denote the estimates $b0$ for $\beta0$ and $b1$ for $\beta1$, then $\hat{y}$ is the predicted or fitted value.

True regression line: $Y = \beta0 + \beta1x$

Estimated or fitted line: $\hat{y} = b0 + b1x$

We expect that the fitted line should be closer to the true regression line when a large amount of data are available.

Residual: Error in Fit

A residual is essentially an error in the fit of the model ˆy = b0 + b1x

Given a set of regression data {(xi, yi); i = 1, 2, . . . , n} and a fitted model, ˆyi =b0 + b1xi, the ith residual ei is given by:

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

$$y_i = b_0 + b_1 x_i + e_i$$

If a set of n residuals is large, then the fit of the model is not good. Small residuals are a sign of a good fit.

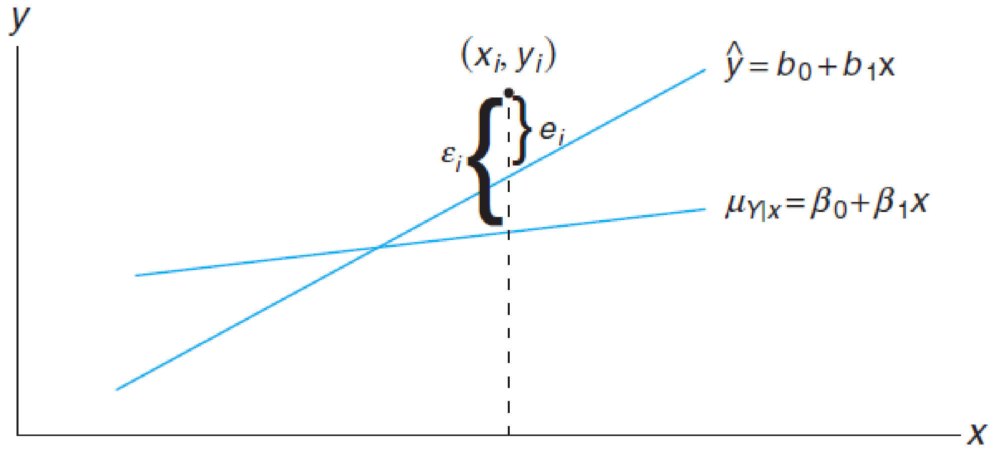


Figure 11.5: Comparing ϵ_i with the residual, e_i.

Least Squares Estimators (LSE)

Least squares: minimization procedure for estimating the parameters. We shall find b0 and b1, the estimates of β0 and β1, so that the sum of the squares of the residuals/errors (SSE) is a minimum. Differentiating SSE with respect to b0,b1.

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

Given the sample {(xi, yi); i = 1, 2, . . . , n}, the least squares estimates b0 and b1 of the regression coefficients β0 and β1 are computed from the formulas

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b_0 = \frac{\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1\bar{x}.$$

Example: fitted line for a real-life pollution study

33 samples of chemically treated waste in a study conducted at Virginia Tech.

x: % reduction in total solids,
y: % reduction in chemical oxygen demand

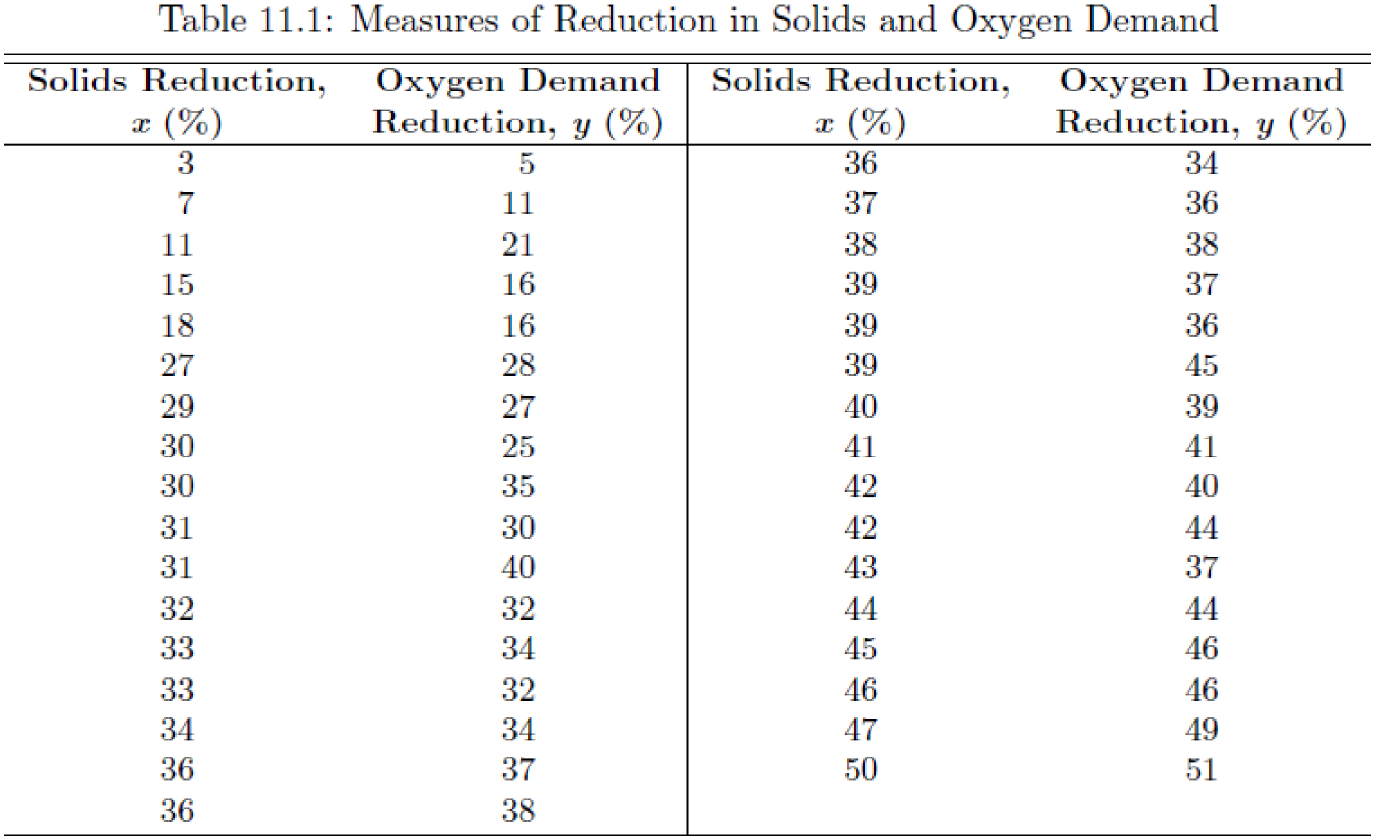


Table 11.1: Measures of Reduction in Solids and Oxygen Demand

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		

Scatter Diagram & Regression Lines

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

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

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

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

the estimated regression line is given by

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

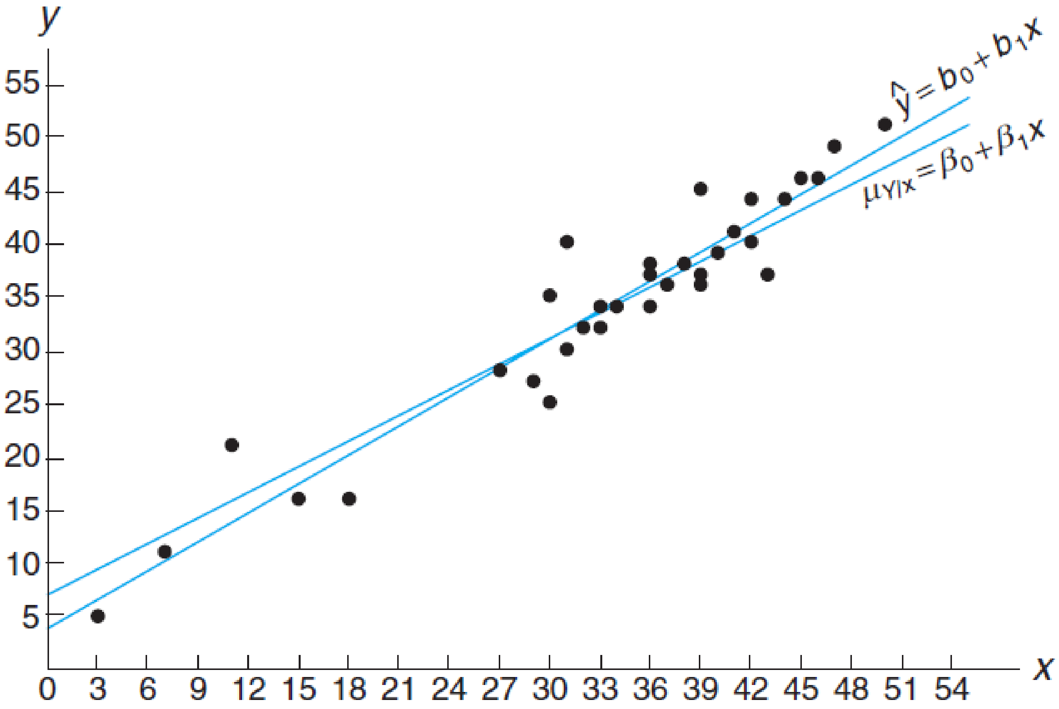


Figure 11.3: Scatter diagram with regression lines.

Properties LSE , Model

$Y = β0 + β1x + \boldsymbol{\varepsilon}$

1. $\sum ei = 0$

2. Ɛi berdistribusi Normal (mean μ= 0 dan variansi σ2)

3. SSE = $\sum(ei)^2$ minimum

4. Taksiran b0 , bi = tak bias

5. Jika X= X rataan maka Ytopi = Y rataan

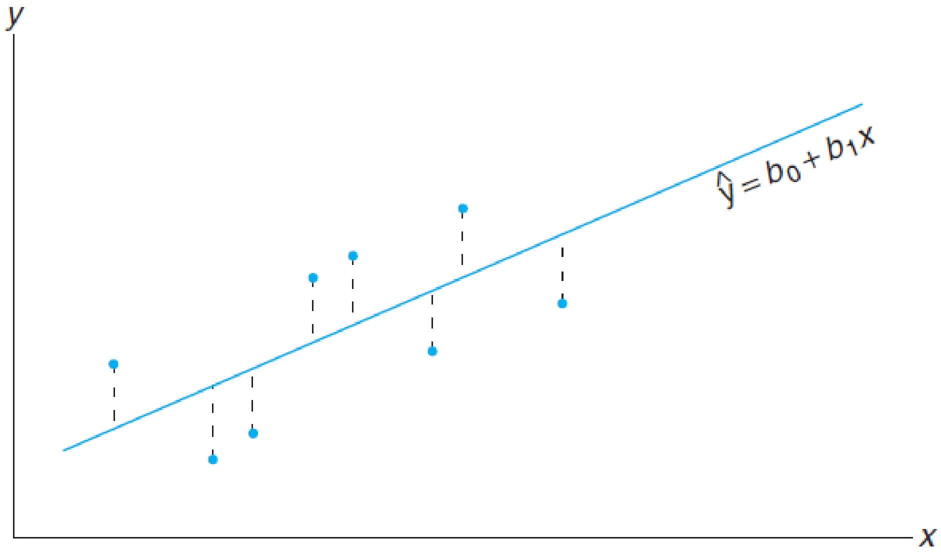


Figure 11.6: Residuals as vertical deviations.

A Measure of Quality of Fit: Coefficient of Determination, R^2

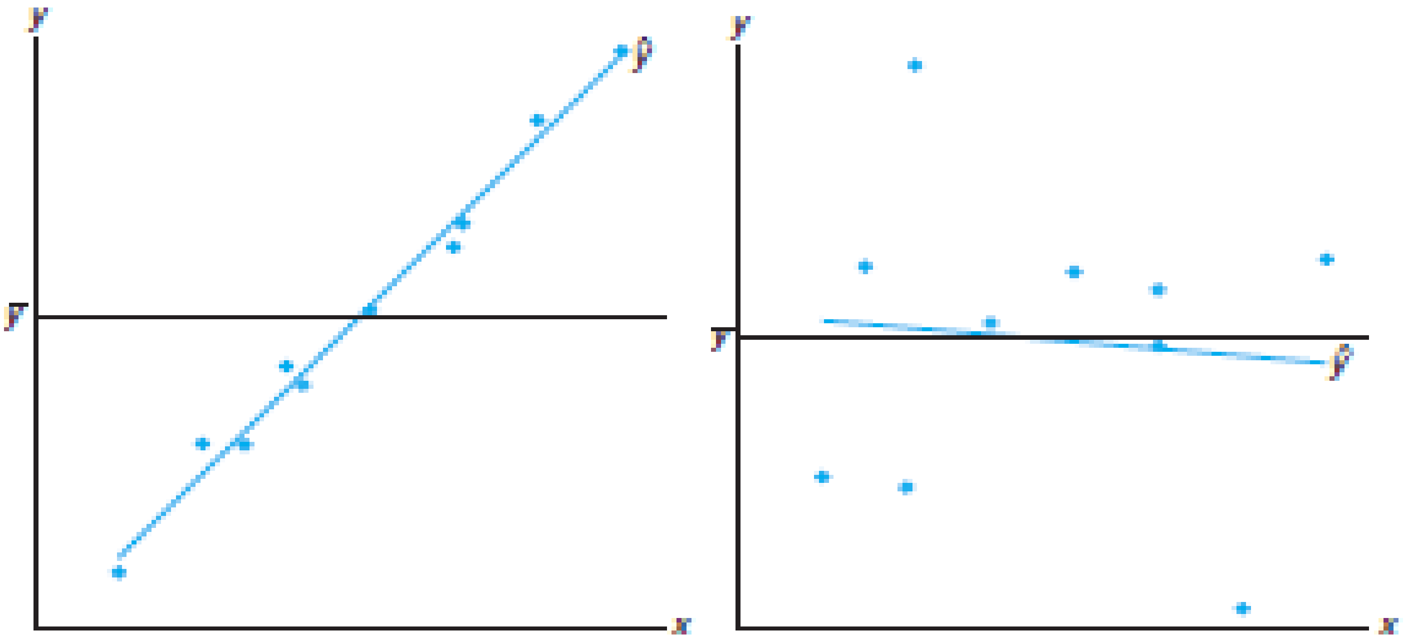
- Besaran, R^2 disebut koefisien determinasi adalah suatu ukuran proposi dari variansi model fitted (regresi) dan variansi variable response.
- Variansi model fitted (regresi)= SSE = Sum Square of Error = $\sum_{i=1}^{n}(y_i - \widehat{y_i})$
- Variansi variable response = SST= Sum Square of Total

= $\sum_{i=1}^{n}(y_i - \overline{y_i})$

- $R^2 = \left(1 - \frac{SSE}{SST}\right)$

A Measure of Quality of Fit: Coefficient of Determination, R^2

- $R^2 = \left(1 - \frac{SSE}{SST}\right)$
- Nilai koefisien determinasi antara 0 dan 1
- Nilai R^2 = 1 artinya garis regresi fit sempurna (perfect), SSE =0
- Nilai Nilai R^2 = 0 artinya garis regresi tidak fit sempurna, SSE =SST (hampir sama)



(a) $R^2 \approx 1.0$

(b) $R^2 \approx 0$

Figure 11.10: Plots depicting a very good fit and a poor fit.

Transformations

- Simple linear regression, both x and y enter the model in a linear fashion.
- Often to work with an alternative model in which either x or y (or both) enters in nonlinear way.
- A transformation of the data may be indicated because a simple plotting of the data may suggest the need to re-express the variable in the model.
- A model in which x or y is transformed should not be viewed as a nonlinear regression model. We normally refer to a regression model as linear

Table 11.6 Some useful Tansformations to Linearize

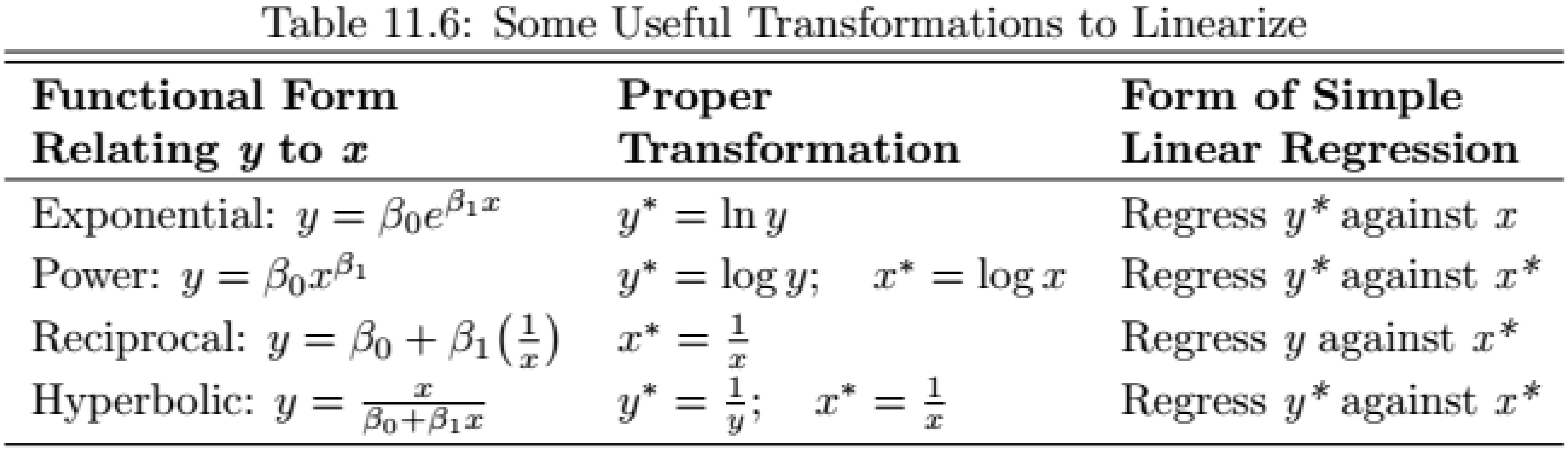


Table 11.6: Some Useful Transformations to Linearize

Functional Form Relating y to x	**Proper Transformation**	**Form of Simple Linear Regression**
Exponential: $y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	Regress y^* against x
Power: $y = \beta_0 x^{\beta_1}$	$y^* = \log y; \quad x^* = \log x$	Regress y^* against x^*
Reciprocal: $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Regress y against x^*
Hyperbolic: $y = \frac{x}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y}; \quad x^* = \frac{1}{x}$	Regress y^* against x^*

Diagram Table 11.6

- Exponential function

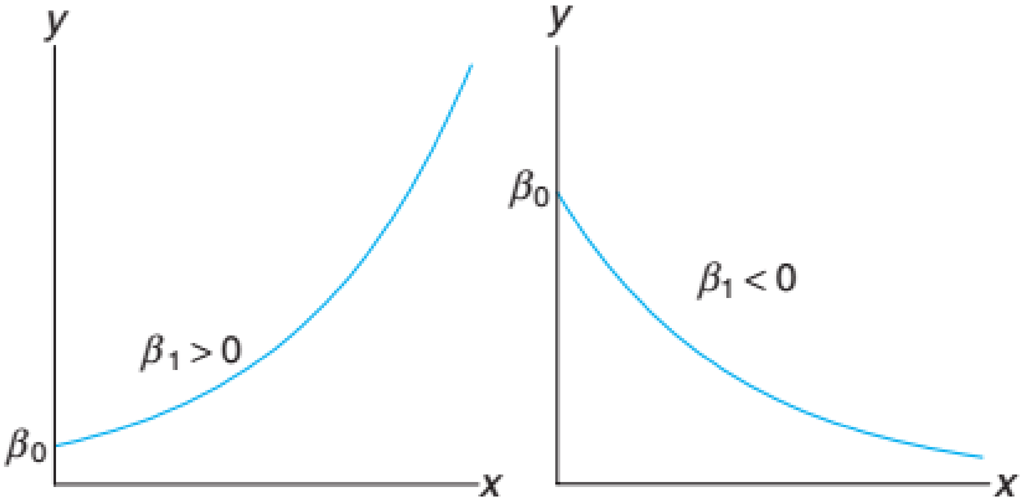


Diagram Table 11.6

- Power function

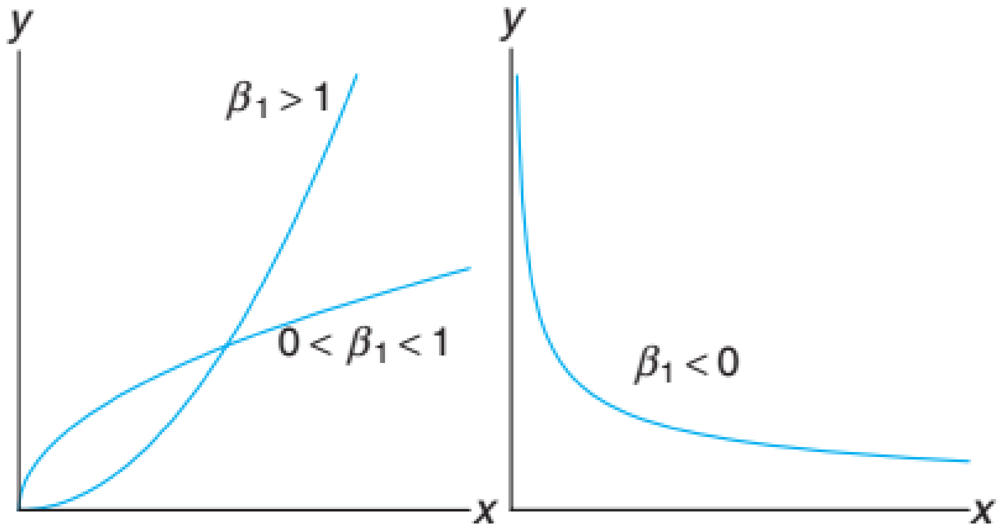
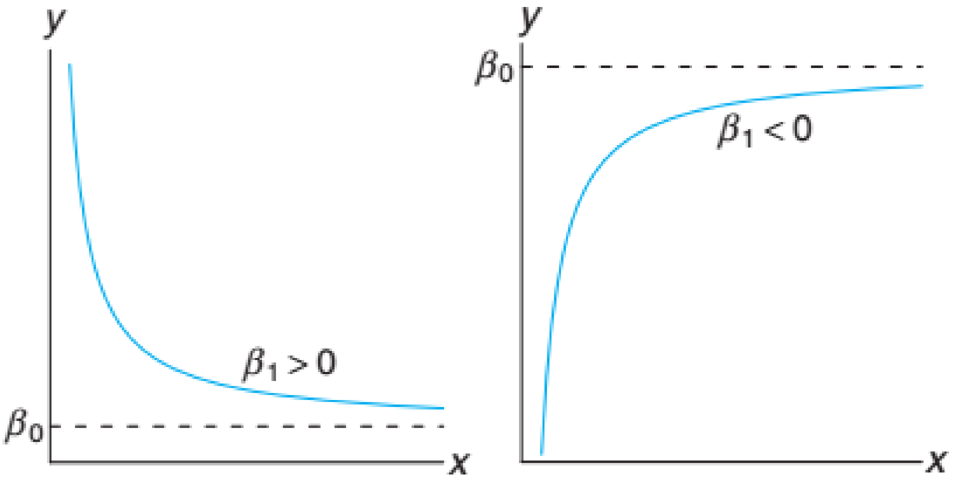


Diagram Table 11.6

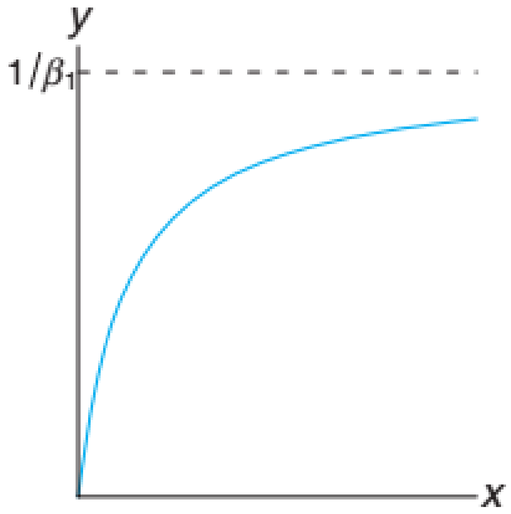
- Reciprocal function



(c) Reciprocal function

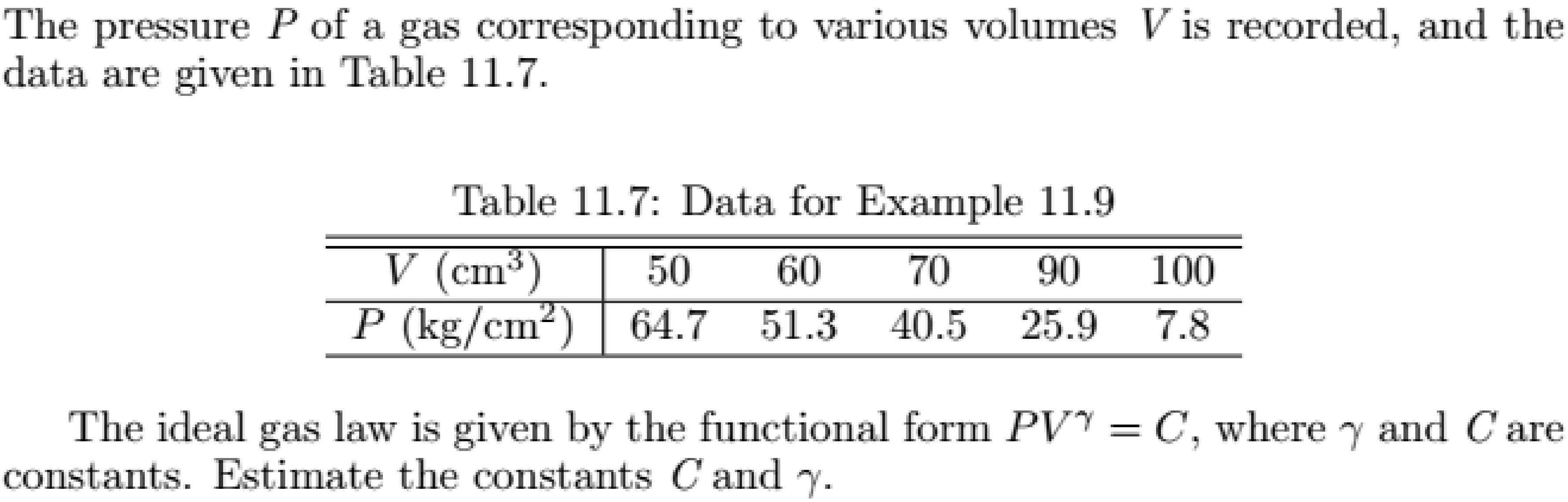
Diagram Table 11.6

- Hyperbolic function



(d) Hyperbolic function

Example 11.9



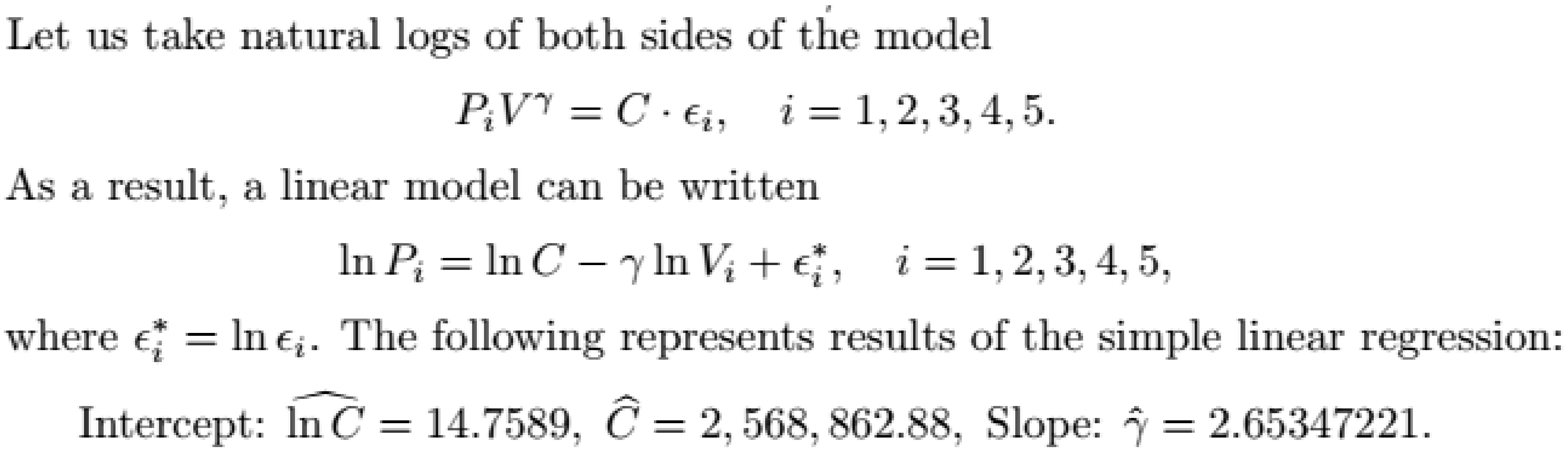
The pressure P of a gas corresponding to various volumes V is recorded, and the data are given in Table 11.7.

Table 11.7: Data for Example 11.9

V (cm^3)	50	60	70	90	100
P (kg/cm^2)	64.7	51.3	40.5	25.9	7.8

The ideal gas law is given by the functional form $PV^\gamma = C$, where γ and C are constants. Estimate the constants C and γ.

Jawab Example 11.9



Let us take natural logs of both sides of the model

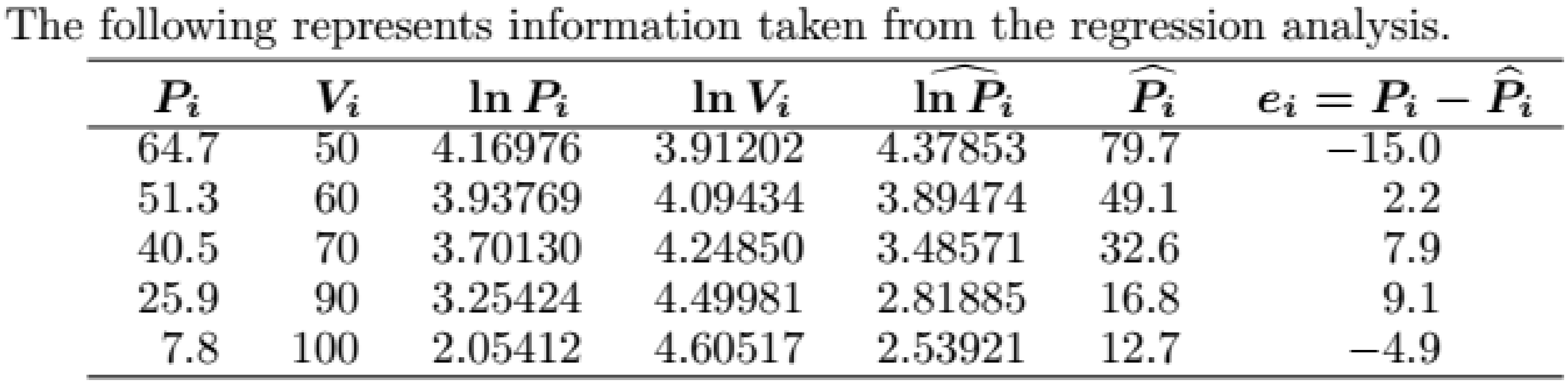
$$P_i V^\gamma = C \cdot \epsilon_i, \quad i = 1, 2, 3, 4, 5.$$

As a result, a linear model can be written

$$\ln P_i = \ln C - \gamma \ln V_i + \epsilon_i^*, \quad i = 1, 2, 3, 4, 5,$$

where $\epsilon_i^* = \ln \epsilon_i$. The following represents results of the simple linear regression:

$$\text{Intercept: } \widehat{\ln C} = 14.7589, \ \widehat{C} = 2,568,862.88, \ \text{Slope: } \hat{\gamma} = 2.65347221.$$



The following represents information taken from the regression analysis.

P_i	V_i	$\ln P_i$	$\ln V_i$	$\widehat{\ln P_i}$	$\widehat{P_i}$	$e_i = P_i - \widehat{P_i}$
64.7	50	4.16976	3.91202	4.37853	79.7	−15.0
51.3	60	3.93769	4.09434	3.89474	49.1	2.2
40.5	70	3.70130	4.24850	3.48571	32.6	7.9
25.9	90	3.25424	4.49981	2.81885	16.8	9.1
7.8	100	2.05412	4.60517	2.53921	12.7	−4.9

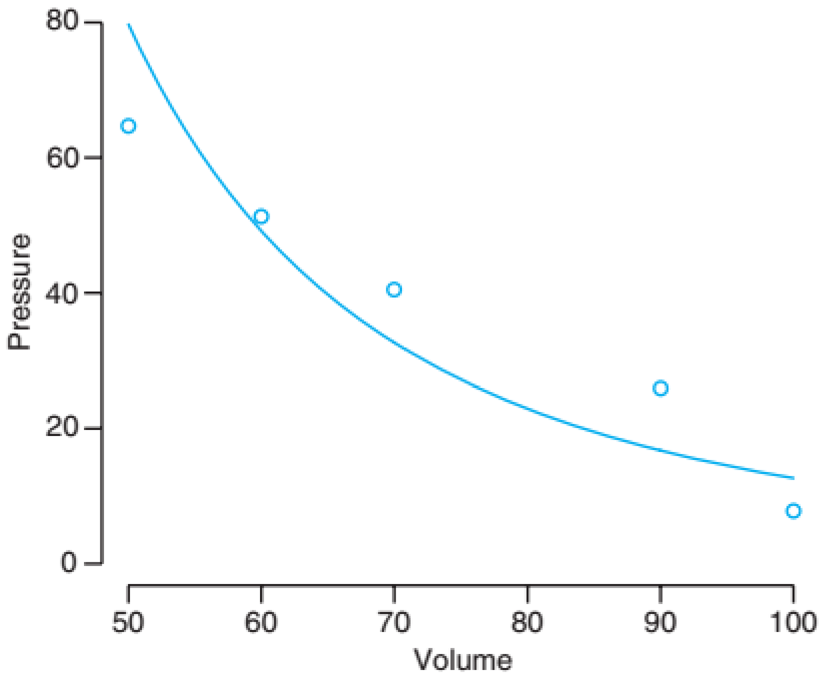
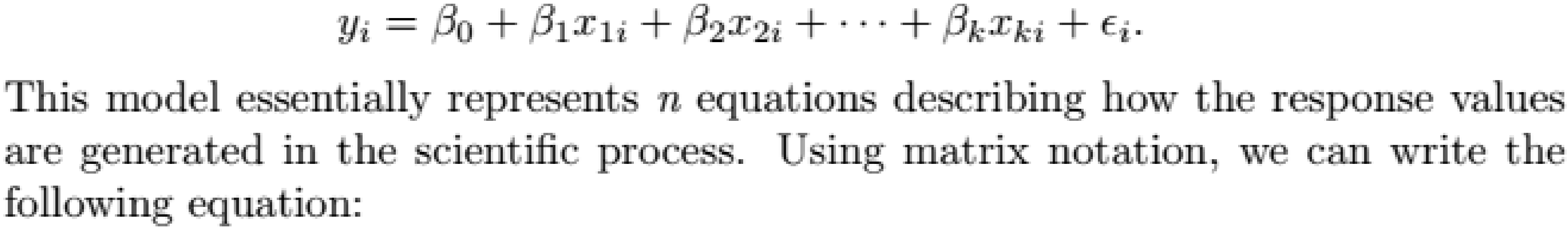


Figure 11.20: Pressure and volume data and fitted regression.

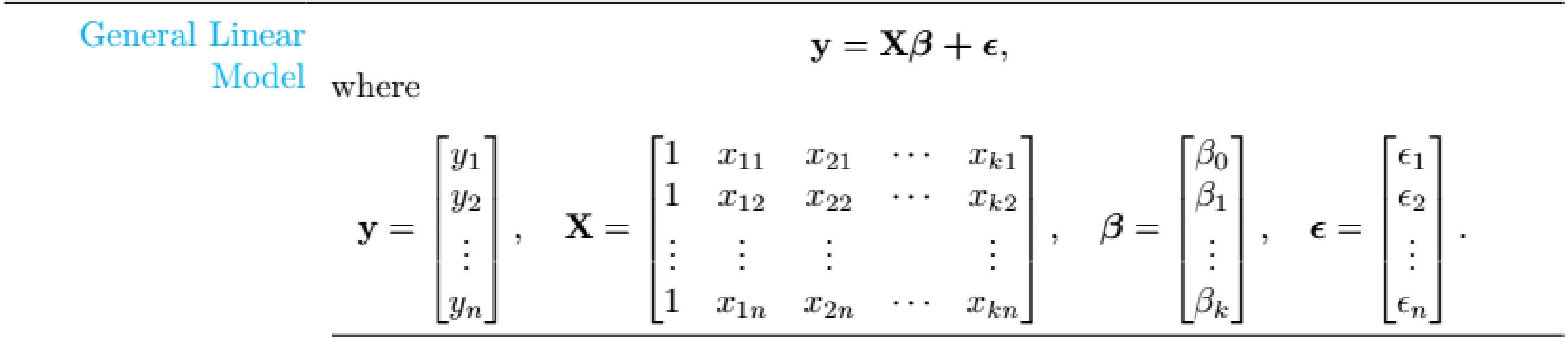
Multiple Linear Regression

Multiple linear Regression. Solution using Geometric Algebra



$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i.$$

This model essentially represents n equations describing how the response values are generated in the scientific process. Using matrix notation, we can write the following equation:



General Linear Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Bab 11.12 Correlation

- Correlation Coefficient attempts to measure the strength of relationship between two variables X, and Y
- Pada Bab 4.2 Kovariansi dua variabel random *X* dan *Y* dengan rataan μ_x dan μ_y adalah

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y$$

- Variabel random X dan Y dengan covariansi σ_{xy} dan simpangan baku masing-masing σ_x dan σ_y. Koefisien korelasi X, Y adalah ρ_{xy}
  - $\rho_{xy} = (\sigma_{xy})/(\sigma_x)(\sigma_y)$
- Nilai koefisien korelasi $-1 \leq \rho_{xy} \leq 1$

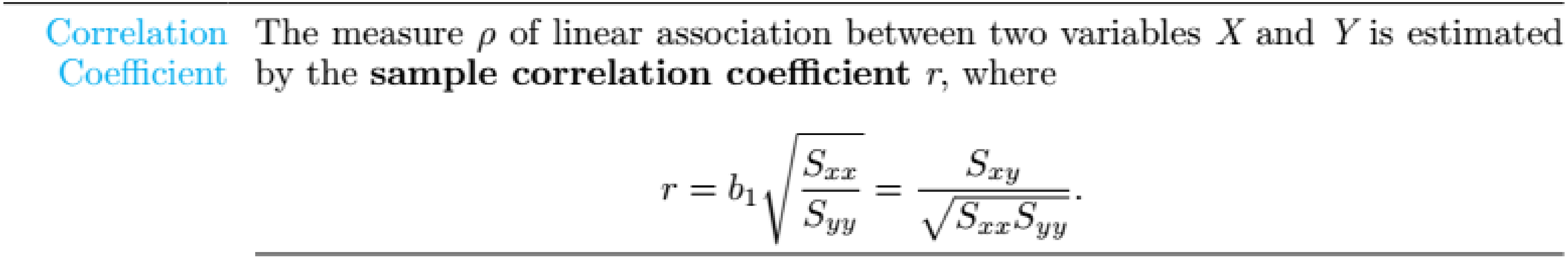
Correlation and Regression

- Regression line $y_i = b_0 + b_1 x_i + e_i$
- The value correlation coefficient ρ is 0 when b1 = 0, which results when there essentially is no linear regression, the regression line is horizontal and any knowledge of X is useless in predict Y.
- The value ρ = 1 if b > 0 and value ρ = -1 if b < 0
- Thus a value of ρ equal to +1 implies a perfect linear relationship with a positive slope while a value of ρ equal to -1 results from a perfect linear relationship with a negative slope.

The Sample Correlation Coefficient

- Menghitung koefisien korelasi

$$
\begin{aligned}
SSE &= \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2b_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx} = S_{yy} - b_1 S_{xy},
\end{aligned}
$$

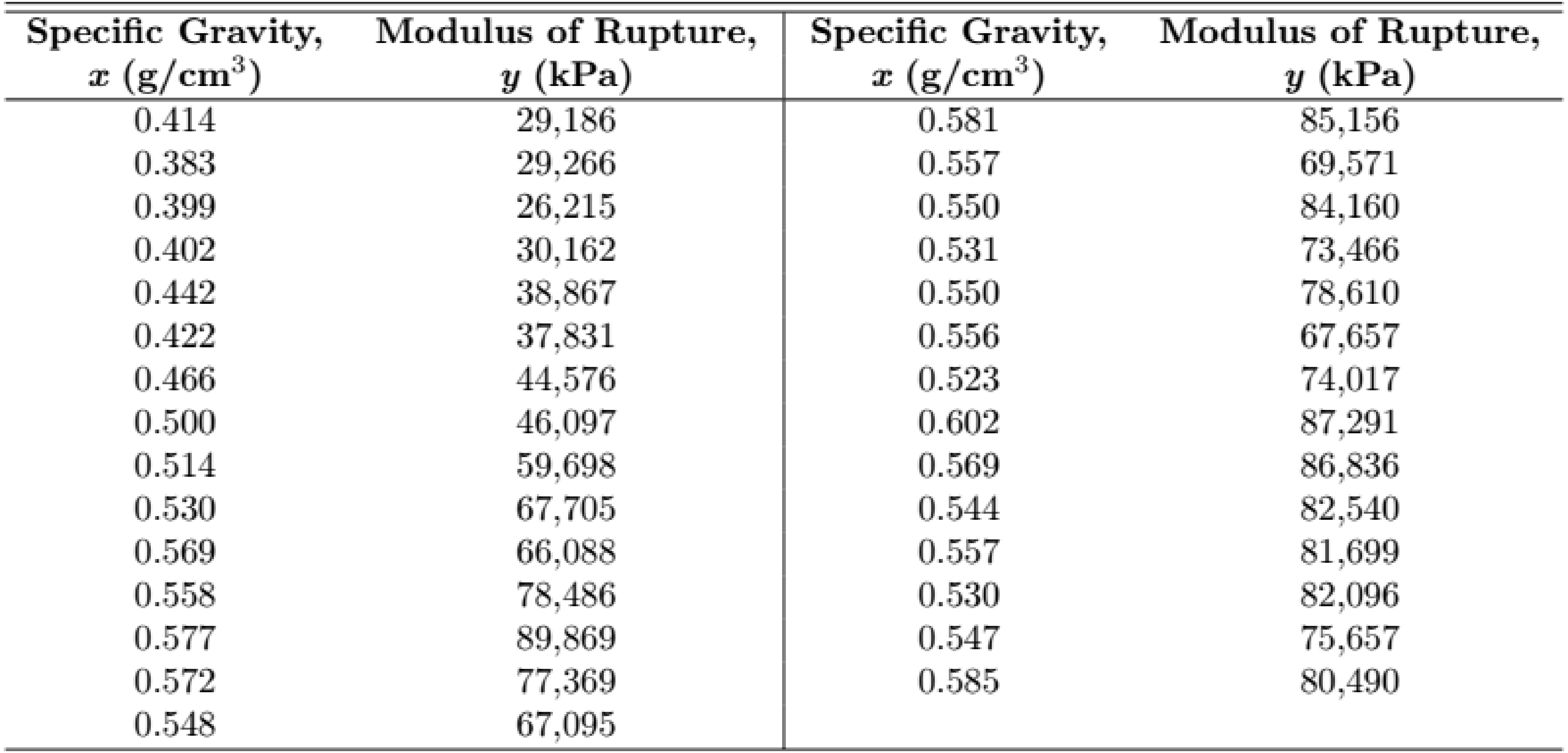


Correlation Coefficient The measure ρ of linear association between two variables X and Y is estimated by the **sample correlation coefficient** r, where

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

$$r = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_1^n (x_i - \bar{x})^2\right)\left(\sum_1^n (y_i - \bar{y})^2\right)}}$$

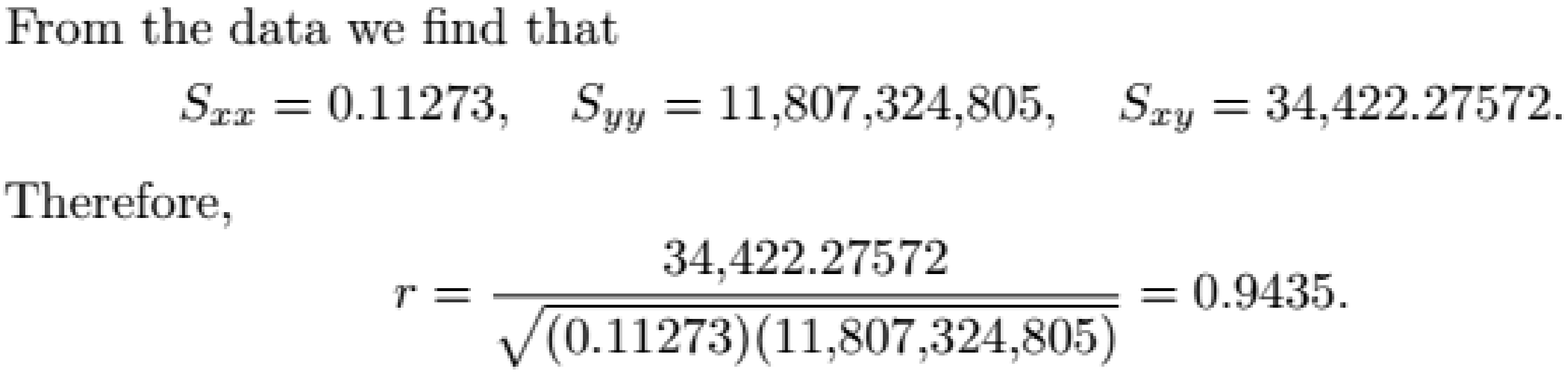
Table 11.9: Data on 29 Loblolly Pines for Example 11.10



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		

Example 11.10

- Table 9 data for example 11.10



From the data we find that

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

Therefore,

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