Dec 25, 2025

Engineering Statistics

Week 10: Simple linear regression

Introduction

We described methods for making inferences about population means. The mean of a population has been treated as a constant, and we have shown how to use sample data to estimate or to test hypotheses about this constant mean.

In many materials engineering applications, however, the mean of a population is not viewed as a constant, but rather as a variable that depends on one or more material or process parameters. For example, the mean tensile strength of a metallic alloy may be treated as a variable that depends on the carbon content of the alloy. A simple relationship might be expressed as

$$\textbf{Mean tensile strength} = 250\ \text{MPa} + 120\ \text{MPa} \times (\textbf{Carbon content in wt\%})$$

Introduction

- Discuss situations in which the mean of the population is treated as a variable, dependent on the value of another variable.
- Present the simplest of all models relating a population mean to another variable: the straight-line model.
- How to use the sample data to estimate the straight-line relationship between the mean value of one variable, y, as it relates to a second variable or variables.
- The methodology of estimating and using a straight-line relationship is referred to as simple linear regression analysis.

Introduction

- In general, regression analysis concerns the study of relationships between quantitative variables to identify, estimate, and validate these relationships.
- The estimated relationship can then be used to predict one variable from the value of the other variable(s).
- The functional relationship between y and x can take two different forms: deterministic or stochastic.

Correlation

Correlation coefficient

The correlation coefficient (r) is a measure of the strength of the linear relation between the two variables. *r takes values between -1 and 1.*

Correlation coefficient

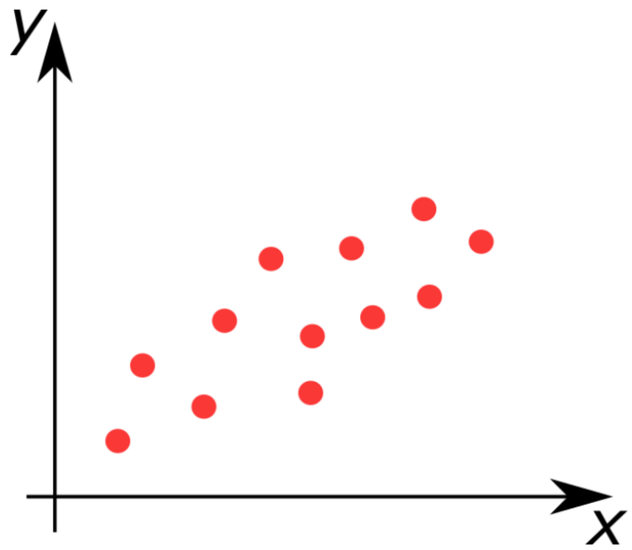
Suppose that we have n pairs of data as follows:

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

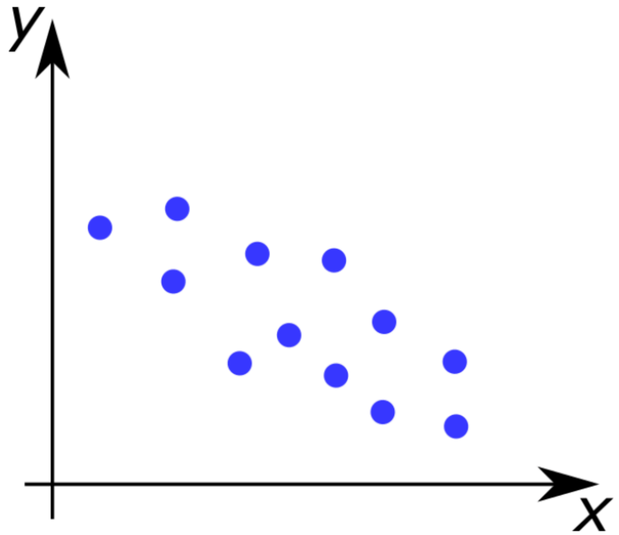
Formula for correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

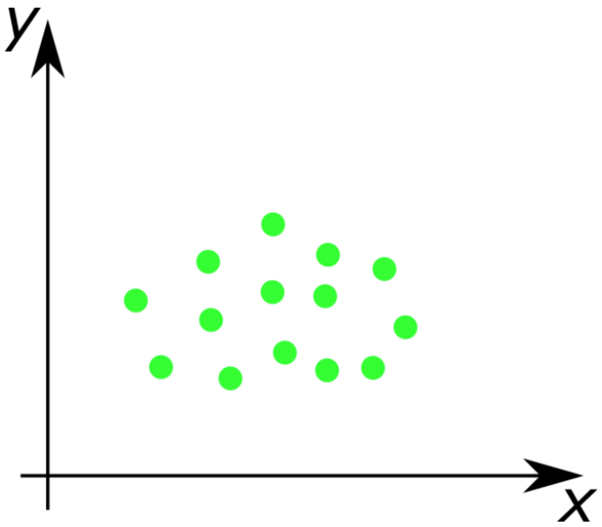
Correlation coefficient



If r is close to 1, there is a strong positive linear relationship between x and y.



If r is close to 1, there is a strong negative linear relationship between x and y.



If r is close to 0, there is no relationship between x and y.

Example

A zoologist collected 20 wild lizards in the southwestern United States. After measuring their total length (mm), they were placed treadmill and their speed (m /sec) recorded.

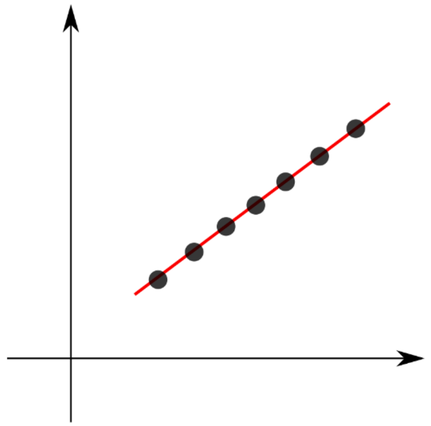
Speed	1.28	1.36	1.24	2.47	1.94	2.52	2.67	1.29	1.56	2.66
Length	179	157	169	146	143	131	159	142	141	130

Calculate the correlation coefficient and comment on it.

Probabilistic models

Deterministic model

If we were to construct a model that hypothesized an exact relationship between variables, it would be called a deterministic model.

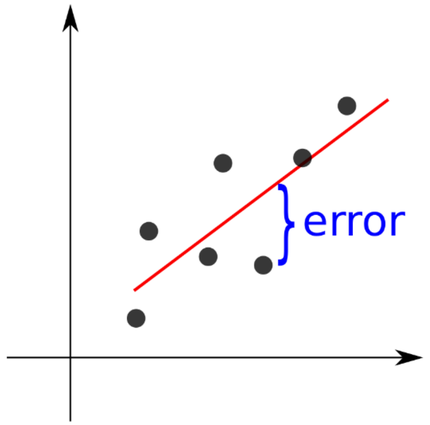


Deterministic model

$$y = f(x)$$

Stochastic model

There will be unexplained variation– perhaps caused by important, but unincluded, variables or by random phenomena – we discard the deterministic model and use a model that accounts for this random error.



Stochastic model

$$y = f(x) + \varepsilon$$

The stochastic model will include both a deterministic component and a random error component.

Simple linear regression model

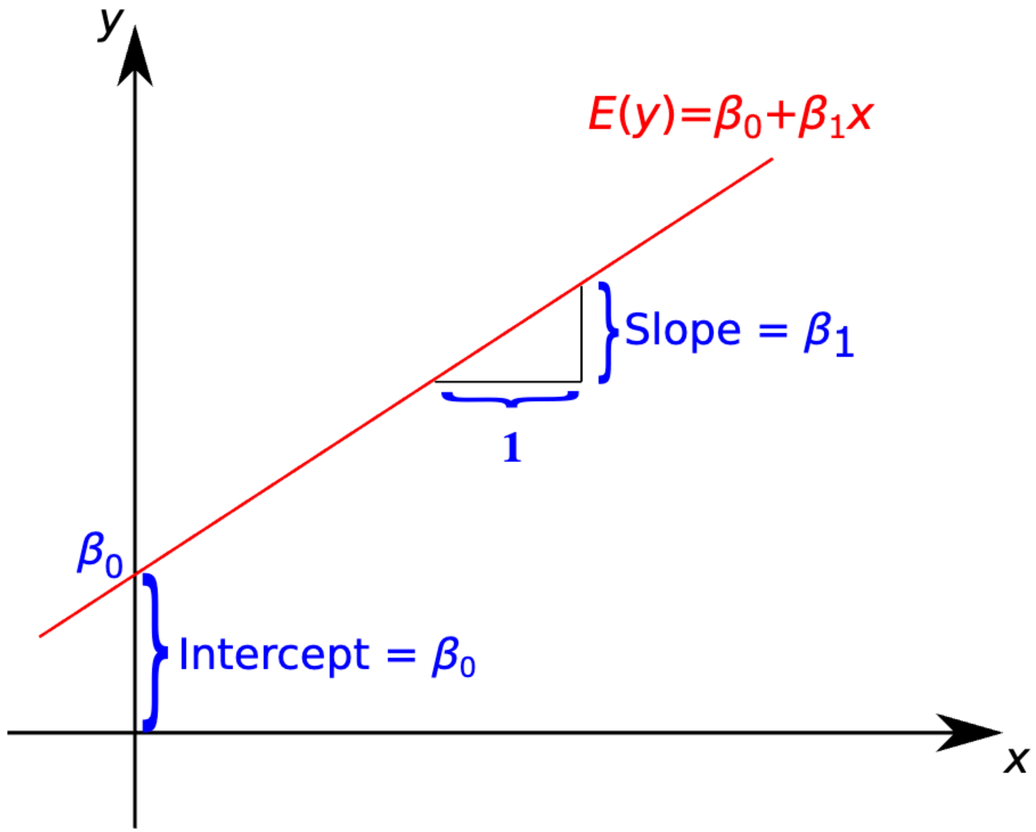
Simple linear regression model

- Suppose that the distribution of ε is normal with mean 0 and constant variance $\sigma 2$, i.e., $\varepsilon \sim N(0, \sigma 2)$.
- Then the mean response at any value of the regressor variable is

  $E(y) = \beta 0 + \beta 1 x.$

Poluation regression model

E (y) = β0 + β1x is called a population regression model.



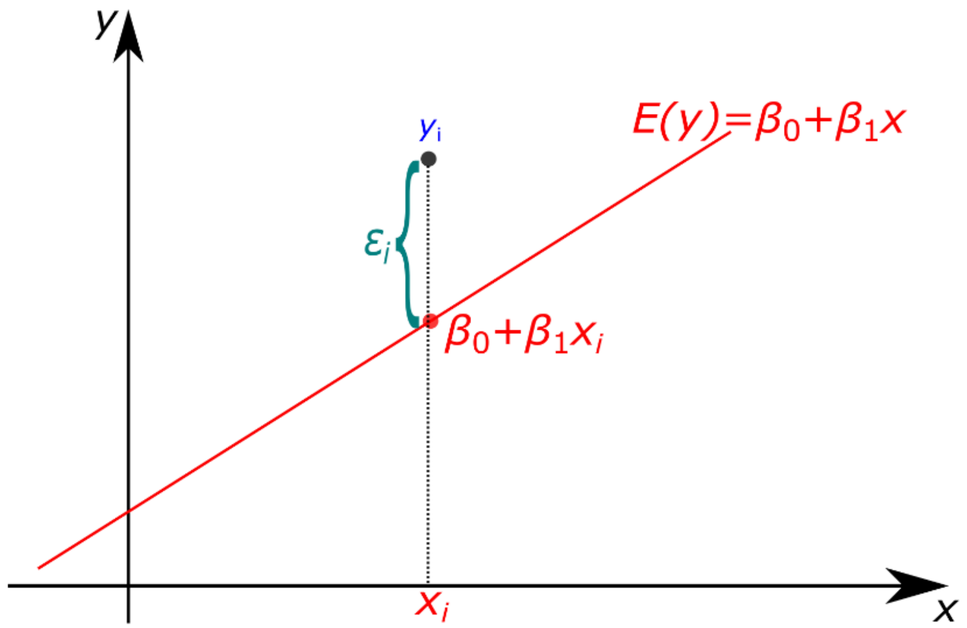
Intercept is the point at which the line intersects or cuts the y–axis.

Slope measures the change (amount of increase or decrease) in the deterministic component of y for every one-unit increase in x.

Fitting the model: The least squares approach

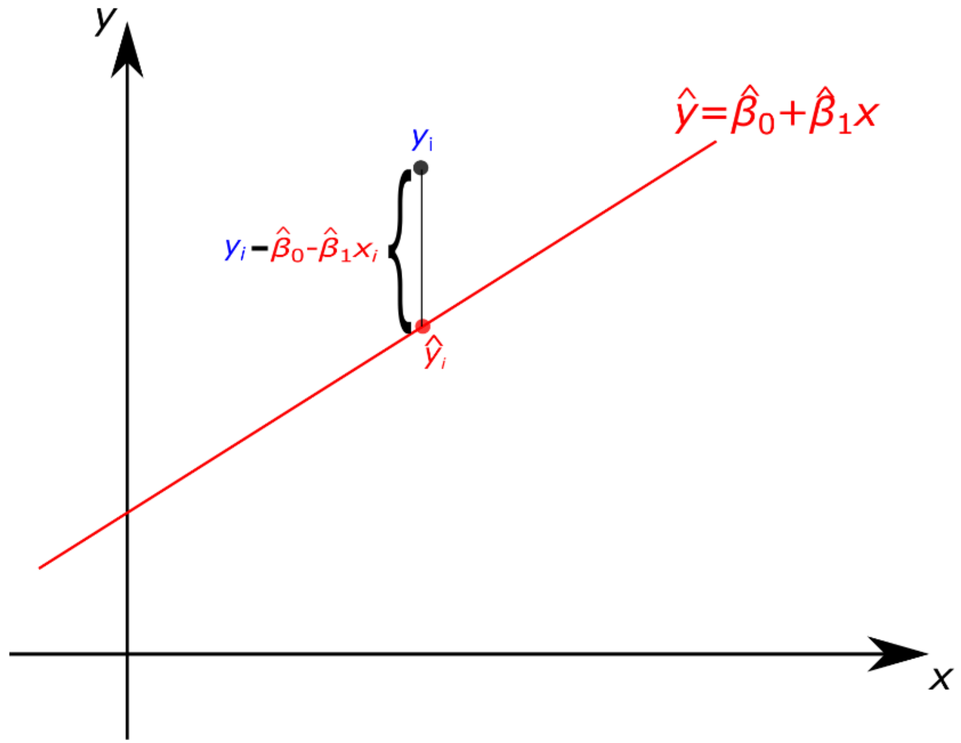
Suppose that we have n pairs of data as given in the following table:

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n



Fitting the model: The least squares approach

Suppose that an arbitrary line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is drawn on the scatter diagram as in the following Figure:



Fitting the model: The least squares approach

Sum of squares

$$S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- The idea underlying least squares (LS) method is minimizing sum of squares of difference between the observations y_i and the straight line ($\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$).

Fitting the model: The least squares approach

After taking partial derivatives of S with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, following equations are obtained:

LS equations

$$\frac{\partial S}{\partial \hat{\beta}_0} = (-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = (-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

Fitting the model: The least squares approach

Making some basic algebraic operations gives **normal equations** as follows:

Normal equations

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

It is clear that $\hat{\beta}_0$ and $\hat{\beta}_1$ are solutions of normal equations.

Fitting the model: The least squares approach

LS estimators

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are the LS estimators of β_0 and β_1, respectively.

Example

In aircraft structures, fuselage skin panels are typically manufactured within a narrow thickness range. An aerospace materials engineer investigates whether the **mean tensile strength** of an aluminum alloy fuselage panel can be approximated as a linear function of the **panel thickness**.

a) Estimate the regression coefficients
b) Interpret the fitted regression model

Panel thickness (mm)	Tensile strength (MPa)
2	298
2.2	301
2.4	304
2.5	306
2.6	309

Example

Step 1: Compute the sample means

$$\bar{x} = \frac{2.0 + 2.2 + 2.4 + 2.5 + 2.6}{5} = 2.34$$

$$\bar{y} = \frac{298 + 301 + 304 + 306 + 309}{5} = 303.6$$

Example

Step 2: Compute deviations and sums

Panel thickness (mm)	Tensile strength (MPa)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
2	298	-0.34	-5.6	1.90	0,116
2.2	301	-0.14	-2.6	0.36	0,020
2.4	304	0.06	0.4	0.02	0,004
2.5	306	0.16	2.4	0.38	0,026
2.6	309	0.26	5.4	1.40	0,068

Example

Step 3: Least squares estimates

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{4.06}{0.234} \sim 17.35$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 303.6 - 17.35 \sim 262.9$$

Example

Step 4: Fitted regression model and interpretation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 262.9 + 17.35x$$

For each **1 mm increase** in panel thickness, the **mean tensile strength** is estimated to increase by approximately **17.35 MPa**.

Course materials

You can download the notes and codes from:

https://github.com/mcavs/ESTUMatse_2022Fall_EngineeringStatistics

Contact

Do not hesitate to contact me on:

https://twitter.com/mustafa__cavus

https://www.linkedin.com/in/mustafacavusphd/

mustafacavus@eskisehir.edu.tr