CEE 260/MIE 273: Probability and Statistics in Civil Engineering

Lecture M7a: Correlation and Variance Analyses in Linear Regression

Jimi Oke

December 4, 2025

Outline

Today's objectives

Today's objectives

- Learn how to compute and interpret the correlation coefficient
- Understand and apply linear regression
- Analyze regression fitness metrics (in particular, R^2)

Historical note: regression analysis

Historical note: regression analysis

The term **regression analysis** was introduced in the 1880s by British statistician, Francis Galton, in his investigation on the relationship between father's height x and son's height y.

Historical note: regression analysis

The term **regression analysis** was introduced in the 1880s by British statistician, Francis Galton, in his investigation on the relationship between father's height x and son's height y.

The trend he observed is called the **regression effect**

Image source: https://www.britannica.com/biography/Francis-Galton

Historical note: regression analysis

The term **regression analysis** was introduced in the 1880s by British statistician, Francis Galton, in his investigation on the relationship between father's height x and son's height y.

The trend he observed is called the **regression effect**

Image source: https://www.britannica.com/biography/Francis-Galton

Historical note: regression analysis

The term **regression analysis** was introduced in the 1880s by British statistician, Francis Galton, in his investigation on the relationship between father's height x and son's height y.

The trend he observed is called the **regression effect**

Image source: https://www.britannica.com/biography/Francis-Galton

- First, he collected paired data (x_i, y_i)

Historical note: regression analysis

The term **regression analysis** was introduced in the 1880s by British statistician, Francis Galton, in his investigation on the relationship between father's height x and son's height y.

The trend he observed is called the **regression effect**

Image source: https://www.britannica.com/biography/Francis-Galton

- First, he collected paired data (x_i, y_i)
- He used the principle of least squares to estimate the regression line

Historical note: regression analysis

The term **regression analysis** was introduced in the 1880s by British statistician, Francis Galton, in his investigation on the relationship between father's height x and son's height y.

The trend he observed is called the **regression effect**

Image source: https://www.britannica.com/biography/Francis-Galton

- First, he collected paired data (x_i, y_i)
- He used the principle of least squares to estimate the regression line
- The goal was to predict son's height from father's height

Historical note: regression analysis

The term **regression analysis** was introduced in the 1880s by British statistician, Francis Galton, in his investigation on the relationship between father's height x and son's height y.

The trend he observed is called the **regression effect**

Image source: https://www.britannica.com/biography/Francis-Galton

- First, he collected paired data (x_i, y_i)
- He used the principle of least squares to estimate the regression line
- The goal was to predict son's height from father's height
- From his results, he found that the height of the son was always *pulled back* ("regressed") toward the mean

Historical note: regression analysis

The term **regression analysis** was introduced in the 1880s by British statistician, Francis Galton, in his investigation on the relationship between father's height x and son's height y.

The trend he observed is called the **regression effect**

Image source: https://www.britannica.com/biography/Francis-Galton

- First, he collected paired data (x_i, y_i)
- He used the principle of least squares to estimate the regression line
- The goal was to predict son's height from father's height
- From his results, he found that the height of the son was always *pulled back* ("regressed") toward the mean
  - E.g. the height of the son of an taller-than-average father was greater than average but not by as much as his father's

Historical note: regression analysis

The term **regression analysis** was introduced in the 1880s by British statistician, Francis Galton, in his investigation on the relationship between father's height x and son's height y.

The trend he observed is called the **regression effect**

Image source: https://www.britannica.com/biography/Francis-Galton

- First, he collected paired data (x_i, y_i)
- He used the principle of least squares to estimate the regression line
- The goal was to predict son's height from father's height
- From his results, he found that the height of the son was always *pulled back* ("regressed") toward the mean
    - E.g. the height of the son of an taller-than-average father was greater than average but not by as much as his father's
    - And the height of the son a shorter-than-average father was lower but not by as much as his father's

Regression analysis

Regression analysis

Regression analysis is used to investigate the relationship between two or more variables.

Regression analysis

Regression analysis is used to investigate the relationship between two or more variables.

Examples of variables with nondeterministic relationships

Regression analysis

Regression analysis is used to investigate the relationship between two or more variables.

Examples of variables with nondeterministic relationships

- x = age of a child; y = size of child's vocabulary
- x = size of an engine; y = fuel efficiency for a car equipped with engine
- x = applied tensile force; y deformation of a metal strip

Regression analysis

Regression analysis is used to investigate the relationship between two or more variables.

Examples of variables with nondeterministic relationships

- $x =$ age of a child; $y =$ size of child's vocabulary
- $x =$ size of an engine; $y =$ fuel efficiency for a car equipped with engine
- $x =$ applied tensile force; y deformation of a metal strip

In this module, we will cover the following key topics:

- Correlation and variance analyses
- Simple Linear Regression and Least Squares Estimation
- Inference for Linear Regression

Covariance

Recall that the **variance** of a random variable X is given by:

Covariance

Recall that the **variance** of a random variable X is given by:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{1}$$

Covariance

Recall that the **variance** of a random variable X is given by:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \quad (1)$$

Then given two r.v.'s X and Y, the **covariance** measures the strength of the **linear** relationship between them.

Covariance

Recall that the **variance** of a random variable X is given by:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{1}$$

Then given two r.v.'s X and Y, the **covariance** measures the strength of the **linear** relationship between them.

Covariance

Covariance

Recall that the **variance** of a random variable X is given by:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{1}$$

Then given two r.v.'s X and Y, the **covariance** measures the strength of the **linear** relationship between them.

Covariance

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{2}$$

Covariance

Recall that the **variance** of a random variable X is given by:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{1}$$

Then given two r.v.'s X and Y, the **covariance** measures the strength of the **linear** relationship between them.

Covariance

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{2}$$

This can also be rewritten as:

Covariance

Recall that the **variance** of a random variable X is given by:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{1}$$

Then given two r.v.'s X and Y, the **covariance** measures the strength of the **linear** relationship between them.

Covariance

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{2}$$

This can also be rewritten as:

$$\text{Cov}(X, Y) = \begin{cases} \sum_x \sum_y (x - \mu_x)(y - \mu_y) f(x, y) & (X, Y) \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_x) f(x, y) dxdy & (X, Y) \text{ continuous} \end{cases} \tag{3}$$

Correlation

Correlation

Correlation

Correlation coefficient

This is the normalized covariance

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{4}$$

Correlation

Correlation coefficient

This is the normalized covariance

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{4}$$

Thus, the correlation coefficient ranges from -1 (perfectly linear negative relationship) to $+1$ (perfectly linear positive relationship).

identify_relationships_lin_neg_strong.pdf

identify_relationships_lin_pos_str

Sample correlation coefficient

Sample correlation coefficient

For a set of n pairs of observations, the **sample correlation coefficient** is given by:

Sample correlation coefficient

For a set of n pairs of observations, the **sample correlation coefficient** is given by:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \tag{5}$$

Sample correlation coefficient

For a set of n pairs of observations, the **sample correlation coefficient** is given by:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \tag{5}$$

where

$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) \tag{6}$$

$$S_{xx} = \sum (x_i - \overline{x})^2 \tag{7}$$

$$S_{yy} = \sum (y_i - \overline{y})^2 \tag{8}$$

Properties of the sample correlation coefficient $\hat{\rho}$

[1] "iff" $\equiv$ "if and only if".

Properties of the sample correlation coefficient $\hat{\rho}$

1. Value does not depend on which of the two variables is labeled x or y

[1] "iff" $\equiv$ "if and only if".

Properties of the sample correlation coefficient $\hat{\rho}$

1. Value does not depend on which of the two variables is labeled x or y
2. Independent of the units in which x and y are measured

[1] "iff" $\equiv$ "if and only if".

Properties of the sample correlation coefficient $\hat{\rho}$

1. Value does not depend on which of the two variables is labeled x or y
2. Independent of the units in which x and y are measured
3. $-1 \leq \hat{\rho} \leq 1$

[1] "iff" $\equiv$ "if and only if".

Properties of the sample correlation coefficient $\hat{\rho}$

1. Value does not depend on which of the two variables is labeled x or y
2. Independent of the units in which x and y are measured
3. $-1 \leq \hat{\rho} \leq 1$
4. $\hat{\rho} = 1$ if and only if all data pairs lie on a straight line with positive slope and $\hat{\rho} = -1$ iff[1] all pairs lie on a straight line with negative slope

[1] "iff" $\equiv$ "if and only if".

Correlation coefficient estimate

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_x \cdot \sigma_y}$$

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Thus, given a **sample** of paired observations X and Y, the estimate $\hat{\rho}$ is:

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Thus, given a **sample** of paired observations X and Y, the estimate $\hat{\rho}$ is:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Thus, given a **sample** of paired observations X and Y, the estimate $\hat{\rho}$ is:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{10}$$

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Thus, given a **sample** of paired observations X and Y, the estimate $\hat{\rho}$ is:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{10}$$

Note that:

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Thus, given a **sample** of paired observations X and Y, the estimate $\hat{\rho}$ is:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{10}$$

Note that:

$$\sum(x_i - \overline{x})(y_i - \overline{y})$$

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Thus, given a **sample** of paired observations X and Y, the estimate $\hat{\rho}$ is:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{10}$$

Note that:

$$\sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i \tag{11}$$

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Thus, given a **sample** of paired observations X and Y, the estimate $\hat{\rho}$ is:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{10}$$

Note that:

$$\sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i \tag{11}$$

Also:

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Thus, given a **sample** of paired observations X and Y, the estimate $\hat{\rho}$ is:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{10}$$

Note that:

$$\sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i \tag{11}$$

Also:

$$\sum(x_i - \overline{x})^2$$

Correlation coefficient estimate

Recall the definition of the correlation coefficient of random variables X and Y:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y))}{\sigma_x \cdot \sigma_y} \tag{9}$$

Thus, given a **sample** of paired observations X and Y, the estimate $\hat{\rho}$ is:

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{10}$$

Note that:

$$\sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i \tag{11}$$

Also:

$$\sum(x_i - \overline{x})^2 = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 \tag{12}$$

Sample correlation coefficient

Sample correlation coefficient

Thus, we obtain the equation for the sample correlation coefficient:

Sample correlation coefficient

Thus, we obtain the equation for the sample correlation coefficient:

$$\hat{\rho} \quad = \quad \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}\sqrt{\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2}}$$

Sample correlation coefficient

Thus, we obtain the equation for the sample correlation coefficient:

$$
\begin{aligned}
\hat{\rho} &= \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}\sqrt{\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2}} \\
&= \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}}
\end{aligned}
$$

Computing the correlation coefficient from summary data

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations

Based on data from the sample concentrations of two pollutants, we are given that

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations

Based on data from the sample concentrations of two pollutants, we are given that

$$
\begin{aligned}
n &= 16 \\
\sum x_i &= 1.656
\end{aligned}
$$

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations

Based on data from the sample concentrations of two pollutants, we are given that

$$\begin{aligned} n &= 16 \\ \sum x_i &= 1.656 \\ \sum y_i &= 170.6 \end{aligned}$$

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations

Based on data from the sample concentrations of two pollutants, we are given that

$$
\begin{aligned}
n &= 16 \\
\sum x_i &= 1.656 \\
\sum y_i &= 170.6 \\
\sum x_i^2 &= 0.196912
\end{aligned}
$$

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations

Based on data from the sample concentrations of two pollutants, we are given that

$$
\begin{aligned}
n &= 16 \\
\sum x_i &= 1.656 \\
\sum y_i &= 170.6 \\
\sum x_i^2 &= 0.196912 \\
\sum x_i y_i &= 20.0397
\end{aligned}
$$

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations

Based on data from the sample concentrations of two pollutants, we are given that

$$\begin{aligned} n &= 16 \\ \sum x_i &= 1.656 \\ \sum y_i &= 170.6 \\ \sum x_i^2 &= 0.196912 \\ \sum x_i y_i &= 20.0397 \\ \sum y_i^2 &= 2253.56 \end{aligned}$$

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations

Based on data from the sample concentrations of two pollutants, we are given that

$$
\begin{aligned}
n &= 16 \\
\sum x_i &= 1.656 \\
\sum y_i &= 170.6 \\
\sum x_i^2 &= 0.196912 \\
\sum x_i y_i &= 20.0397 \\
\sum y_i^2 &= 2253.56
\end{aligned}
$$

Find the sample correlation coefficient.

Computing the correlation coefficient from summary data

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations (cont.)

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations (cont.)

$$\hat{\rho} =$$

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations (cont.)

$$\hat{\rho} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sqrt{\sum x_i^2 - \frac{1}{n} \left(\sum x_i\right)^2} \sqrt{\sum y_i^2 - \frac{1}{n} \left(\sum y_i\right)^2}}$$

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations (cont.)

$$\hat{\rho} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}\sqrt{\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2}}$$

$$= \frac{20.0397 - \frac{1}{16}(1.656)(170.6)}{\sqrt{0.196912 - \frac{1}{16}(1.656)^2}\sqrt{20.0397 - \frac{1}{16}(170.6)^2}}$$

Computing the correlation coefficient from summary data

Example 1: Pollutant concentrations (cont.)

$$
\begin{aligned}
\hat{\rho} &= \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}\sqrt{\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2}} \\
&= \frac{20.0397 - \frac{1}{16}(1.656)(170.6)}{\sqrt{0.196912 - \frac{1}{16}(1.656)^2}\sqrt{20.0397 - \frac{1}{16}(170.6)^2}} \\
&= \boxed{0.716}
\end{aligned}
$$

Simple linear regression model

Simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

Simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{13}$$

Simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{13}$$

random error term

Simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{13}$$

where ϵ is a normally distributed random variable:

random error term

Simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{13}$$

where ϵ is a normally distributed random variable:

random error term

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{14}$$

Simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{13}$$

where ϵ is a normally distributed random variable:

random error term

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{14}$$

β_0 (intercept) and β_1 (slope) are the **regression coefficients**

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$,

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, the random error term ϵ allows (x_i, y_i) to fall:

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, the random error term ϵ allows (x_i, y_i) to fall:

- above the line:

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, the random error term ϵ allows (x_i, y_i) to fall:

- above the line:

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, the random error term ϵ allows (x_i, y_i) to fall:

- above the line: $\epsilon > 0$
- below the line:

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, the random error term ϵ allows (x_i, y_i) to fall:

- above the line: $\epsilon > 0$
- below the line:

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, the random error term ϵ allows (x_i, y_i) to fall:

- above the line: $\epsilon > 0$
- below the line: $\epsilon < 0$
- on the line:

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, the random error term ϵ allows (x_i, y_i) to fall:

- above the line: $\epsilon > 0$
- below the line: $\epsilon < 0$
- on the line:

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, the random error term ϵ allows (x_i, y_i) to fall:

- above the line: $\epsilon > 0$
- below the line: $\epsilon < 0$
- on the line: $\epsilon = 0$

Error term

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, the random error term ϵ allows (x_i, y_i) to fall:

- above the line: $\epsilon > 0$
- below the line: $\epsilon < 0$
- on the line: $\epsilon = 0$

Image source: https://sigmazone.com/labrea_scatter_plots/

The observed errors in model predictions are known as **residuals**.

Alternative notation

Alternative notation

Define

$$\mu_{Y \cdot x} = \mathbb{E}(Y|X = x) = \text{expected (or mean) value of } Y \text{ when } X = x$$
$$\sigma^2_{Y \cdot x} = \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y \cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y \cdot x}$ describes the variability of y values when $X = x$.

Example: Age and vocabulary size of children

Let:

$$x = \text{age of a child}$$
$$y = \text{vocabulary size}$$

Then $\mu_{Y \cdot 5}$ is the average vocabulary size for all 5-year-old children in the population.
And $\sigma^2_{Y \cdot 5}$ indicates the amount of variability in vocabulary size for 5-year-olds.

Alternative notation

Define

$$\mu_{Y\cdot x} = \mathbb{E}(Y|X = x)$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y\cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y\cdot x}$ describes the variability of y values when $X = x$.

Example: Age and vocabulary size of children

Let:

$$x = \text{age of a child}$$
$$y = \text{vocabulary size}$$

Then $\mu_{Y\cdot 5}$ is the average vocabulary size for all 5-year-old children in the population.
And $\sigma^2_{Y\cdot 5}$ indicates the amount of variability in vocabulary size for 5-year-olds.

Alternative notation

Define

$$\mu_{Y \cdot x} = \mathbb{E}(Y|X = x)$$

$$= \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y \cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y \cdot x}$ describes the variability of y values when $X = x$.

Example: Age and vocabulary size of children

Let:

$$x = \text{age of a child}$$

$$y = \text{vocabulary size}$$

Then $\mu_{Y \cdot 5}$ is the average vocabulary size for all 5-year-old children in the population.
And $\sigma^2_{Y \cdot 5}$ indicates the amount of variability in vocabulary size for 5-year-olds.

Alternative notation

Define

$$
\begin{aligned}
\mu_{Y\cdot x} &= \mathbb{E}(Y|X = x) = \text{expected (or mean) value of } Y \text{ when } X = x \\
\sigma^2_{Y\cdot x} &= \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x
\end{aligned}
$$

Alternative notation

Define

$$\mu_{Y\cdot x} = \mathbb{E}(Y|X = x) = \text{expected (or mean) value of } Y \text{ when } X = x$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y\cdot x}$ is the mean of all y values for which $X = x$

Alternative notation

Define

$$\mu_{Y \cdot x} = \mathbb{E}(Y|X = x) = \text{expected (or mean) value of } Y \text{ when } X = x$$
$$\sigma^2_{Y \cdot x} = \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y \cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y \cdot x}$ describes the variability of y values when $X = x$.

Alternative notation

Define

$$\mu_{Y \cdot x} = \mathbb{E}(Y|X = x) = \text{expected (or mean) value of } Y \text{ when } X = x$$
$$\sigma^2_{Y \cdot x} = \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y \cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y \cdot x}$ describes the variability of y values when $X = x$.

Example: Age and vocabulary size of children

Alternative notation

Define

$$\mu_{Y\cdot x} = \mathbb{E}(Y|X=x) = \text{expected (or mean) value of } Y \text{ when } X = x$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(Y|X=x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y\cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y\cdot x}$ describes the variability of y values when $X = x$.

Example: Age and vocabulary size of children

Let:

Alternative notation

Define

$$\mu_{Y\cdot x} = \mathbb{E}(Y|X = x) = \text{expected (or mean) value of } Y \text{ when } X = x$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y\cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y\cdot x}$ describes the variability of y values when $X = x$.

Example: Age and vocabulary size of children

Let:

$$x = \text{age of a child}$$
$$y = \text{vocabulary size}$$

Then $\mu_{Y\cdot 5}$ is the average vocabulary size for all 5-year-old children in the population.
And $\sigma^2_{Y\cdot 5}$ indicates the amount of variability in vocabulary size for 5-year-olds.

Alternative notation

Define

$$\mu_{Y\cdot x} = \mathbb{E}(Y|X = x) = \text{expected (or mean) value of } Y \text{ when } X = x$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y\cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y\cdot x}$ describes the variability of y values when $X = x$.

Example: Age and vocabulary size of children

Let:

$$x = \text{age of a child}$$
$$y = \text{vocabulary size}$$

Alternative notation

Define

$$\mu_{Y\cdot x} = \mathbb{E}(Y|X = x) = \text{expected (or mean) value of } Y \text{ when } X = x$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y\cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y\cdot x}$ describes the variability of y values when $X = x$.

Example: Age and vocabulary size of children

Let:

$$x = \text{age of a child}$$
$$y = \text{vocabulary size}$$

Then $\mu_{Y\cdot 5}$ is the average vocabulary size for all 5-year-old children in the population.

Alternative notation

Define

$$\mu_{Y \cdot x} = \mathbb{E}(Y|X = x) = \text{expected (or mean) value of } Y \text{ when } X = x$$
$$\sigma^2_{Y \cdot x} = \mathbb{V}(Y|X = x) = \text{variance of } Y \text{ when } X = x$$

Thus, $\mu_{Y \cdot x}$ is the mean of all y values for which $X = x$ and $\sigma^2_{Y \cdot x}$ describes the variability of y values when $X = x$.

Example: Age and vocabulary size of children

Let:

$$x = \text{age of a child}$$
$$y = \text{vocabulary size}$$

Then $\mu_{Y \cdot 5}$ is the average vocabulary size for all 5-year-old children in the population.
And $\sigma^2_{Y \cdot 5}$ indicates the amount of variability in vocabulary size for 5-year-olds.

Further discussions on regression equation

Further discussions on regression equation

We see that:

Further discussions on regression equation

We see that:

$$\mu_{Y\cdot x} = \mathbb{E}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \beta_0 + \beta_1 x + \mathbb{E}(\boldsymbol{\epsilon}) = \beta_0 + \beta_1 x + 0$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\boldsymbol{\epsilon}) = 0 + \sigma^2$$

Implications:

Further discussions on regression equation

We see that:

$$\mu_{Y\cdot x} = \beta_0 + \beta_1 x + \mathbb{E}(\epsilon) = \beta_0 + \beta_1 x + 0$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \epsilon) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\epsilon) = 0 + \sigma^2$$

Implications:

Further discussions on regression equation

We see that:

$$\begin{aligned}\mu_{Y\cdot x} &= \beta_0 + \beta_1 x + 0 \\ \sigma^2_{Y\cdot x} &= \mathbb{V}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\boldsymbol{\epsilon}) = 0 + \sigma^2\end{aligned}$$

Implications:

Further discussions on regression equation

We see that:

$$\mu_{Y\cdot x} =$$

$$\sigma^2_{Y\cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\boldsymbol{\epsilon}) = 0 + \sigma^2$$

Implications:

Further discussions on regression equation

We see that:

$$\mu_{Y \cdot x} =$$

$$\mathbb{V}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\boldsymbol{\epsilon}) = 0 + \sigma^2$$

Implications:

Further discussions on regression equation

We see that:

$$\mu_{Y \cdot x} =$$

$$\mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\epsilon) = 0 + \sigma^2$$

Implications:

Further discussions on regression equation

We see that:

$$\mu_{Y \cdot x} =$$

$$0 + \sigma^2$$

Implications:

Further discussions on regression equation

We see that:

$$\mu_{Y\cdot x} = \mathbb{E}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \beta_0 + \beta_1 x + \mathbb{E}(\boldsymbol{\epsilon}) = \beta_0 + \beta_1 x + 0$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\boldsymbol{\epsilon}) = 0 + \sigma^2$$

Implications:

Further discussions on regression equation

We see that:

$$\mu_{Y \cdot x} = \mathbb{E}(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + \mathbb{E}(\epsilon) = \beta_0 + \beta_1 x + 0$$
$$\sigma^2_{Y \cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \epsilon) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\epsilon) = 0 + \sigma^2$$

Implications:

- The *mean* value of Y is a linear function of x

Further discussions on regression equation

We see that:

$$\mu_{Y\cdot x} = \mathbb{E}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \beta_0 + \beta_1 x + \mathbb{E}(\boldsymbol{\epsilon}) = \beta_0 + \beta_1 x + 0$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\boldsymbol{\epsilon}) = 0 + \sigma^2$$

Implications:

- The *mean* value of Y is a linear function of x
- The true regression line $y = \beta_0 + \beta_1 x$ is the *line of mean values*

Further discussions on regression equation

We see that:

$$\mu_{Y\cdot x} = \mathbb{E}(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + \mathbb{E}(\epsilon) = \beta_0 + \beta_1 x + 0$$
$$\sigma^2_{Y\cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \epsilon) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\epsilon) = 0 + \sigma^2$$

Implications:

- The *mean* value of Y is a linear function of x
- The true regression line $y = \beta_0 + \beta_1 x$ is the *line of mean values*
- Slope β_1 is the expected change in Y for a unit increase in x

Further discussions on regression equation

We see that:

$$\mu_{Y \cdot x} = \mathbb{E}(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + \mathbb{E}(\epsilon) = \beta_0 + \beta_1 x + 0$$
$$\sigma^2_{Y \cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \epsilon) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\epsilon) = 0 + \sigma^2$$

Implications:

- The *mean* value of Y is a linear function of x
- The true regression line $y = \beta_0 + \beta_1 x$ is the *line of mean values*
- Slope β_1 is the expected change in Y for a unit increase in x
- The amount of variability in Y values is the same for each value of x

Further discussions on regression equation

We see that:

$$\mu_{Y \cdot x} = \mathbb{E}(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + \mathbb{E}(\epsilon) = \beta_0 + \beta_1 x + 0$$
$$\sigma^2_{Y \cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \epsilon) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\epsilon) = 0 + \sigma^2$$

Implications:

- The *mean* value of Y is a linear function of x
- The true regression line $y = \beta_0 + \beta_1 x$ is the *line of mean values*
- Slope β_1 is the expected change in Y for a unit increase in x
- The amount of variability in Y values is the same for each value of x

Further discussions on regression equation

We see that:

$$\mu_{Y \cdot x} = \mathbb{E}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \beta_0 + \beta_1 x + \mathbb{E}(\boldsymbol{\epsilon}) = \beta_0 + \beta_1 x + 0$$
$$\sigma^2_{Y \cdot x} = \mathbb{V}(\beta_0 + \beta_1 x + \boldsymbol{\epsilon}) = \mathbb{V}(\beta_0 + \beta_1 x) + \mathbb{V}(\boldsymbol{\epsilon}) = 0 + \sigma^2$$

Implications:

- The *mean* value of Y is a linear function of x
- The true regression line $y = \beta_0 + \beta_1 x$ is the *line of mean values*
- Slope β_1 is the expected change in Y for a unit increase in x
- The amount of variability in Y values is the same for each value of x (**homogeneity of variance**)

Analyzing a regression equation

Example 2: Flow rate

Example 2: Flow rate

The flow rate y (m^3/min) in a device used for air quality measurement depends on the pressure drop x (in. of water) across the device's filter. Suppose that for x values between 5 and 20, the variables are related by the regression model:

Analyzing a regression equation

Example 2: Flow rate

The flow rate y (m^3/min) in a device used for air quality measurement depends on the pressure drop x (in. of water) across the device's filter. Suppose that for x values between 5 and 20, the variables are related by the regression model:

$$y = -0.12 + 0.095x$$

Analyzing a regression equation

Example 2: Flow rate

The flow rate y (m^3/min) in a device used for air quality measurement depends on the pressure drop x (in. of water) across the device's filter. Suppose that for x values between 5 and 20, the variables are related by the regression model:

$$y = -0.12 + 0.095x$$

and $\sigma = 0.025$.

Analyzing a regression equation

Example 2: Flow rate

The flow rate y (m^3/min) in a device used for air quality measurement depends on the pressure drop x (in. of water) across the device's filter. Suppose that for x values between 5 and 20, the variables are related by the regression model:

$$y = -0.12 + 0.095x$$

and $\sigma = 0.025$. Answer the following questions:

(a) What is the expected change in flow rate associated with a 1-inch increase in pressure drop?

Analyzing a regression equation

Example 2: Flow rate

The flow rate y (m^3/min) in a device used for air quality measurement depends on the pressure drop x (in. of water) across the device's filter. Suppose that for x values between 5 and 20, the variables are related by the regression model:

$$y = -0.12 + 0.095x$$

and $\sigma = 0.025$. Answer the following questions:

(a) What is the expected change in flow rate associated with a 1-inch increase in pressure drop?

(b) What is the mean change in flow rate when the pressure drop decreases by 5 in?

Analyzing a regression equation

Example 2: Flow rate

The flow rate y (m^3/min) in a device used for air quality measurement depends on the pressure drop x (in. of water) across the device's filter. Suppose that for x values between 5 and 20, the variables are related by the regression model:

$$y = -0.12 + 0.095x$$

and $\sigma = 0.025$. Answer the following questions:

(a) What is the expected change in flow rate associated with a 1-inch increase in pressure drop?

(b) What is the mean change in flow rate when the pressure drop decreases by 5 in?

(c) What is the expected flow rate for a pressure drop of 10 in?

Analyzing a regression equation

Example 2: Flow rate

The flow rate y (m^3/min) in a device used for air quality measurement depends on the pressure drop x (in. of water) across the device's filter. Suppose that for x values between 5 and 20, the variables are related by the regression model:

$$y = -0.12 + 0.095x$$

and $\sigma = 0.025$. Answer the following questions:

(a) What is the expected change in flow rate associated with a 1-inch increase in pressure drop?

(b) What is the mean change in flow rate when the pressure drop decreases by 5 in?

(c) What is the expected flow rate for a pressure drop of 10 in?

(d) For a pressure drop of 10 in., what is the probability that the observed flow rate will exceed 0.835?

Analyzing a regression equation

Example 2: Flow rate (cont.)

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation:

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and $\beta_1 = 0.095$

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and $\beta_1 = 0.095$

(a) The **expected change** in flow rate associated with a 1-inch increase in pressure drop is the slope:

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and $\beta_1 = 0.095$

(a) The **expected change** in flow rate associated with a 1-inch increase in pressure drop is the slope: 0.095 m^3/min.

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and $\beta_1 = 0.095$

(a) The **expected change** in flow rate associated with a 1-inch increase in pressure drop is the slope: $0.095\ \text{m}^3/\text{min}$.

(b) The **mean change** in flow rate when the pressure drop decreases by 5 in. is given by:

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and $\beta_1 = 0.095$

(a) The **expected change** in flow rate associated with a 1-inch increase in pressure drop is the slope: 0.095 m^3/min.

(b) The **mean change** in flow rate when the pressure drop decreases by 5 in. is given by:

$$0.095(-5) = -0.475 \text{ m}^3/\text{min}$$

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and $\beta_1 = 0.095$

(a) The **expected change** in flow rate associated with a 1-inch increase in pressure drop is the slope: 0.095 m^3/min.

(b) The **mean change** in flow rate when the pressure drop decreases by 5 in. is given by:

$$0.095(-5) = -0.475 \text{ m}^3/\text{min}$$

(c) The expected flow rate for a pressure drop of 10 in. is given by:

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and $\beta_1 = 0.095$

(a) The **expected change** in flow rate associated with a 1-inch increase in pressure drop is the slope: 0.095 m^3/min.

(b) The **mean change** in flow rate when the pressure drop decreases by 5 in. is given by:

$$0.095(-5) = -0.475 \text{ m}^3/\text{min}$$

(c) The expected flow rate for a pressure drop of 10 in. is given by:

$$y = -0.12 + 0.095(10)$$

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and $\beta_1 = 0.095$

(a) The **expected change** in flow rate associated with a 1-inch increase in pressure drop is the slope: 0.095 m^3/min.

(b) The **mean change** in flow rate when the pressure drop decreases by 5 in. is given by:

$$0.095(-5) = -0.475 \text{ m}^3/\text{min}$$

(c) The expected flow rate for a pressure drop of 10 in. is given by:

$$y = -0.12 + 0.095(10) = -0.12 + 0.95$$

Analyzing a regression equation

Example 2: Flow rate (cont.)

From the model equation: $\beta_0 = -0.12$ and $\beta_1 = 0.095$

(a) The **expected change** in flow rate associated with a 1-inch increase in pressure drop is the slope: 0.095 m^3/min.

(b) The **mean change** in flow rate when the pressure drop decreases by 5 in. is given by:

$$0.095(-5) = -0.475 \text{ m}^3/\text{min}$$

(c) The expected flow rate for a pressure drop of 10 in. is given by:

$$y = -0.12 + 0.095(10) = -0.12 + 0.95 = 0.83$$

Analyzing a regression equation

Example 2: Flow rate (cont.)

Analyzing a regression equation

Example 2: Flow rate (cont.)

Analyzing a regression equation

Example 2: Flow rate (cont.)

(d) For $x = 10$, Y has a mean value

Analyzing a regression equation

Example 2: Flow rate (cont.)

(d) For $x = 10$, Y has a mean value $\mu_{Y\cdot 10} = \mathbb{E}(Y|X = 10) = 0.83$

Example 2: Flow rate (cont.)

(d) For $x = 10$, Y has a mean value $\mu_{Y\cdot 10} = \mathbb{E}(Y|X = 10) = 0.83$ (from part (c)).

Example 2: Flow rate (cont.)

(d) For $x = 10$, Y has a mean value $\mu_{Y \cdot 10} = \mathbb{E}(Y|X = 10) = 0.83$ (from part (c)).
Thus:

$$P(Y > 0.835|X = 10) = 1 - P(Y \leq 0.835|X = 10)$$

Analyzing a regression equation

Example 2: Flow rate (cont.)

(d) For $x = 10$, Y has a mean value $\mu_{Y \cdot 10} = \mathbb{E}(Y|X = 10) = 0.83$ (from part (c)).
Thus:

$$
\begin{aligned}
P(Y > 0.835|X = 10) &= 1 - P(Y \leq 0.835|X = 10) \\
&= 1 - \Phi\left(\frac{y - \mu}{\sigma}\right)
\end{aligned}
$$

Analyzing a regression equation

Example 2: Flow rate (cont.)

(d) For $x = 10$, Y has a mean value $\mu_{Y\cdot 10} = \mathbb{E}(Y|X = 10) = 0.83$ (from part (c)).
Thus:

$$\begin{aligned} P(Y > 0.835|X = 10) &= 1 - P(Y \leq 0.835|X = 10) \\ &= 1 - \Phi\left(\frac{y - \mu}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{0.835 - 0.83}{0.025}\right) \end{aligned}$$

Example 2: Flow rate (cont.)

(d) For $x = 10$, Y has a mean value $\mu_{Y\cdot 10} = \mathbb{E}(Y|X = 10) = 0.83$ (from part (c)).
Thus:

$$\begin{aligned} P(Y > 0.835|X = 10) &= 1 - P(Y \leq 0.835|X = 10) \\ &= 1 - \Phi\left(\frac{y - \mu}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{0.835 - 0.83}{0.025}\right) \\ &= 1 - \Phi(0.2) \end{aligned}$$

Analyzing a regression equation

Example 2: Flow rate (cont.)

(d) For $x = 10$, Y has a mean value $\mu_{Y\cdot 10} = \mathbb{E}(Y|X = 10) = 0.83$ (from part (c)).
Thus:

$$\begin{aligned} P(Y > 0.835|X = 10) &= 1 - P(Y \leq 0.835|X = 10) \\ &= 1 - \Phi\left(\frac{y - \mu}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{0.835 - 0.83}{0.025}\right) \\ &= 1 - \Phi(0.2) \\ &= 1 - 0.5793 \end{aligned}$$

Analyzing a regression equation

Example 2: Flow rate (cont.)

(d) For $x = 10$, Y has a mean value $\mu_{Y \cdot 10} = \mathbb{E}(Y|X = 10) = 0.83$ (from part (c)).
Thus:

$$\begin{aligned} P(Y > 0.835|X = 10) &= 1 - P(Y \leq 0.835|X = 10) \\ &= 1 - \Phi\left(\frac{y - \mu}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{0.835 - 0.83}{0.025}\right) \\ &= 1 - \Phi(0.2) \\ &= 1 - 0.5793 \\ &= \boxed{0.4207} \end{aligned}$$

Principle of least squares

Principle of least squares

The *true* values of β_0, β_1 and σ^2 are almost never known.

Principle of least squares

The *true* values of β_0, β_1 and σ^2 are almost never known.
But we can estimate the parameters from sample data

Principle of least squares

The *true* values of β_0, β_1 and σ^2 are almost never known.
But we can estimate the parameters from sample data based on the principle of least squares.

Principle of least squares

The *true* values of β_0, β_1 and σ^2 are almost never known.
But we can estimate the parameters from sample data based on the principle of least squares.

Principle of least squares

Principle of least squares

The *true* values of β_0, β_1 and σ^2 are almost never known.
But we can estimate the parameters from sample data based on the principle of least squares.

Principle of least squares

Given the sum of squared deviations between the sample observations and a candidate regression line as:

Principle of least squares

The *true* values of β_0, β_1 and σ^2 are almost never known.
But we can estimate the parameters from sample data based on the principle of least squares.

Principle of least squares

Given the sum of squared deviations between the sample observations and a candidate regression line as:

$$\Delta^2 = \sum_n [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{15}$$

Then minimizing Δ^2 yields the estimates of the regression coefficients:

Principle of least squares

The *true* values of β_0, β_1 and σ^2 are almost never known.
But we can estimate the parameters from sample data based on the principle of least squares.

Principle of least squares

Given the sum of squared deviations between the sample observations and a candidate regression line as:

$$\Delta^2 = \sum_n [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{15}$$

Then minimizing Δ^2 yields the estimates of the regression coefficients:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{16}$$

Principle of least squares

The *true* values of β_0, β_1 and σ^2 are almost never known.
But we can estimate the parameters from sample data based on the principle of least squares.

Principle of least squares

Given the sum of squared deviations between the sample observations and a candidate regression line as:

$$\Delta^2 = \sum_n [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{15}$$

Then minimizing Δ^2 yields the estimates of the regression coefficients:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{16}$$

$$\hat{\beta}_1 = \frac{\sum_n (x_i - \overline{x})(y_i - \overline{y})}{\sum_n (x_i - \overline{x})^2} \tag{17}$$

Least squares regression in MATLAB

Example 3: Relationship between population and number of accidents

Least squares regression in MATLAB

Example 3: Relationship between population and number of accidents

Using the `accidents` dataset in MATLAB, perform a least-squares regression of accidents in a state *on* the population of the state[a]:

`load accidents`

`x = hwydata(:,14);` (Population of state)

`y = hwydata(:,4);` (Accidents per state)

(a) What are the slope and intercept estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$?

(b) How can you evaluate the strength of the relationship?

[a] See the file `ex3_l19_least_squares_regression.m`

Fitted values and residuals

Fitted values and residuals

Given the estimated regression equation:

Fitted values and residuals

Given the estimated regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{18}$$

The **fitted** (or predicted) **values** $\hat{y}_i$ are obtained by substituting x_i into the regression equation.

Fitted values and residuals

Given the estimated regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{18}$$

The **fitted** (or predicted) **values** $\hat{y}_i$ are obtained by substituting x_i into the regression equation.
The **residuals** are the vertical deviations $y_i - \hat{y}_i$ from the estimated line.

Fitted values and residuals

Given the estimated regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{18}$$

The **fitted** (or predicted) **values** $\hat{y}_i$ are obtained by substituting x_i into the regression equation.
The **residuals** are the vertical deviations $y_i - \hat{y}_i$ from the estimated line.

Summary

Fitted values and residuals

Given the estimated regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{18}$$

The **fitted** (or predicted) **values** $\hat{y}_i$ are obtained by substituting x_i into the regression equation.
The **residuals** are the vertical deviations $y_i - \hat{y}_i$ from the estimated line.

Summary

- Data (Observed) = Fit (Expected) + Residual

Fitted values and residuals

Given the estimated regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{18}$$

The **fitted** (or predicted) **values** $\hat{y}_i$ are obtained by substituting x_i into the regression equation.
The **residuals** are the vertical deviations $y_i - \hat{y}_i$ from the estimated line.

Summary

- Data (Observed) = Fit (Expected) + Residual
- **R**esidual = **O**bserved − **E**xpected

Residuals illustrated

Residuals illustrated

The figure below shows linear models and corersponding residual plots.

Residuals illustrated

The figure below shows linear models and corersponding residual plots.

What patterns do you observe?

Error sum of squares

Error sum of squares

In simple linear regression, the **error sum of squares** (or **residual sum of squares**) is given by:

Error sum of squares

In simple linear regression, the **error sum of squares** (or **residual sum of squares**) is given by:

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{19}$$

Since we assume that the residuals are normally distributed with a constant variance σ^2

Error sum of squares

In simple linear regression, the **error sum of squares** (or **residual sum of squares**) is given by:

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{19}$$

Since we assume that the residuals are normally distributed with a constant variance σ^2 (a property known as **homoskedasticity**)

Error sum of squares

In simple linear regression, the **error sum of squares** (or **residual sum of squares**) is given by:

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{19}$$

Since we assume that the residuals are normally distributed with a constant variance σ^2 (a property known as **homoskedasticity**)
we can estimate σ^2 as:

Error sum of squares

In simple linear regression, the **error sum of squares** (or **residual sum of squares**) is given by:

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{19}$$

Since we assume that the residuals are normally distributed with a constant variance σ^2 (a property known as **homoskedasticity**)
we can estimate σ^2 as:

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} \tag{20}$$

Error sum of squares

In simple linear regression, the **error sum of squares** (or **residual sum of squares**) is given by:

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{19}$$

Since we assume that the residuals are normally distributed with a constant variance σ^2 (a property known as **homoskedasticity**)
we can estimate σ^2 as:

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} \tag{20}$$

The degrees of freedom $df = n - 2$ because 2 parameters must first be estimated before computing $\hat{\sigma}^2$: β_0 and β_1

Total sum of squares

Total sum of squares

The error sum of squares (SSE) measures how much variation in y is *unexplained* by the estimated model.

Total sum of squares

The error sum of squares (SSE) measures how much variation in y is *unexplained* by the estimated model.
The total amount of variation in the observed y values, however, is measured by the

Total sum of squares

The error sum of squares (SSE) measures how much variation in y is *unexplained* by the estimated model.

The total amount of variation in the observed y values, however, is measured by the **total sum of squares**:

Total sum of squares

The error sum of squares (SSE) measures how much variation in y is *unexplained* by the estimated model.

The total amount of variation in the observed y values, however, is measured by the **total sum of squares**:

$$SST = \sum (y_i - \overline{y})^2 \tag{21}$$

Total sum of squares

The error sum of squares (SSE) measures how much variation in y is *unexplained* by the estimated model.
The total amount of variation in the observed y values, however, is measured by the **total sum of squares**:

$$SST = \sum (y_i - \overline{y})^2 \tag{21}$$

Notes

Total sum of squares

The error sum of squares (SSE) measures how much variation in y is *unexplained* by the estimated model.
The total amount of variation in the observed y values, however, is measured by the **total sum of squares**:

$$SST = \sum (y_i - \overline{y})^2 \tag{21}$$

Notes

- The SSE is the sum of squared deviations about the least squares line

Total sum of squares

The error sum of squares (SSE) measures how much variation in y is *unexplained* by the estimated model.
The total amount of variation in the observed y values, however, is measured by the **total sum of squares**:

$$SST = \sum (y_i - \overline{y})^2 \tag{21}$$

Notes

- The SSE is the sum of squared deviations about the least squares line
- The SST is the sum of squared deviations about the horizontal line at height $\overline{y}$

Total sum of squares

The error sum of squares (SSE) measures how much variation in y is *unexplained* by the estimated model.
The total amount of variation in the observed y values, however, is measured by the **total sum of squares**:

$$SST = \sum (y_i - \overline{y})^2 \tag{21}$$

Notes

- The SSE is the sum of squared deviations about the least squares line
- The SST is the sum of squared deviations about the horizontal line at height $\overline{y}$
- $SSE \leq SST$

Total sum of squares

The error sum of squares (SSE) measures how much variation in y is *unexplained* by the estimated model.
The total amount of variation in the observed y values, however, is measured by the **total sum of squares**:

$$SST = \sum (y_i - \overline{y})^2 \tag{21}$$

Notes

- The SSE is the sum of squared deviations about the least squares line
- The SST is the sum of squared deviations about the horizontal line at height $\overline{y}$
- $SSE \leq SST$
- The ratio SSE/SST is the proportion of total variation unexplained by the simple linear regression model

Goodness of linear fit: coefficient of determination

Goodness of linear fit: coefficient of determination

Question

Goodness of linear fit: coefficient of determination

Question

If SSE/SST is the proportion of variance unexplained by the regression model, what is the proportion of variance *explained* by the model?

Goodness of linear fit: coefficient of determination

Question

If SSE/SST is the proportion of variance unexplained by the regression model, what is the proportion of variance *explained* by the model?

Answer: $\boxed{1 - SSE/SST}$

Goodness of linear fit: coefficient of determination

Question

If SSE/SST is the proportion of variance unexplained by the regression model, what is the proportion of variance *explained* by the model?

Answer: $\boxed{1 - SSE/SST}$

To evaluate how well an estimated regression line fits the given data, we use the measure R^2,

Goodness of linear fit: coefficient of determination

Question

If SSE/SST is the proportion of variance unexplained by the regression model, what is the proportion of variance *explained* by the model?

Answer: $\boxed{1 - SSE/SST}$

To evaluate how well an estimated regression line fits the given data, we use the measure R^2, called the **coefficient of determination**.

$$R^2 = 1 - \frac{SSE}{SST} \tag{22}$$

Coefficient of determination R^2

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

then we can rewrite R^2 as:

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

then we can rewrite R^2 as:

$$R^2 = \frac{SSR}{SST} \tag{24}$$

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

then we can rewrite R^2 as:

$$R^2 = \frac{SSR}{SST} \tag{24}$$

Notes

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

then we can rewrite R^2 as:

$$R^2 = \frac{SSR}{SST} \tag{24}$$

Notes

- R^2 is a number between 0 and 1:

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

then we can rewrite R^2 as:

$$R^2 = \frac{SSR}{SST} \tag{24}$$

Notes

- R^2 is a number between 0 and 1:

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

then we can rewrite R^2 as:

$$R^2 = \frac{SSR}{SST} \tag{24}$$

Notes

- R^2 is a number between 0 and 1: $0 \leq R^2 \leq 1$
- $R^2 = 1$ indicates a perfect linear fit (all variance in data is explained by linear model)

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

then we can rewrite R^2 as:

$$R^2 = \frac{SSR}{SST} \tag{24}$$

Notes

- R^2 is a number between 0 and 1: $0 \leq R^2 \leq 1$
- $R^2 = 1$ indicates a perfect linear fit (all variance in data is explained by linear model)
- As R^2 decreases

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

then we can rewrite R^2 as:

$$R^2 = \frac{SSR}{SST} \tag{24}$$

Notes

- R^2 is a number between 0 and 1: $0 \leq R^2 \leq 1$
- $R^2 = 1$ indicates a perfect linear fit (all variance in data is explained by linear model)
- As R^2 decreases

Coefficient of determination R^2

If we define the **regression sum of squares** (SSR) as:

$$SSR = SST - SSE \tag{23}$$

then we can rewrite R^2 as:

$$R^2 = \frac{SSR}{SST} \tag{24}$$

Notes

- R^2 is a number between 0 and 1: $0 \leq R^2 \leq 1$
- $R^2 = 1$ indicates a perfect linear fit (all variance in data is explained by linear model)
- As R^2 decreases $\rightarrow$ a weaker linear fit
- $\hat{\rho}^2$ approximates R^2 for large n

Summary of variance measures

Summary of variance measures

$$\textbf{Error sum of squares} \quad SSE = \sum (y_i - \hat{y}_i)^2 \tag{25}$$

$$= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i \tag{26}$$

Summary of variance measures

$$\textbf{Error sum of squares} \quad SSE = \sum (y_i - \hat{y}_i)^2 \qquad (25)$$

$$= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i \qquad (26)$$

$$\textbf{Total sum of squares} \quad SST = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2 \qquad (27)$$

Summary of variance measures

$$\textbf{Error sum of squares} \quad SSE = \sum (y_i - \hat{y}_i)^2 \tag{25}$$

$$= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i \tag{26}$$

$$\textbf{Total sum of squares} \quad SST = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2 \tag{27}$$

$$\textbf{Estimated variance} \quad \hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = s^2 \tag{28}$$

Summary of variance measures

$$\begin{aligned}
\textbf{Error sum of squares} \quad SSE &= \sum (y_i - \hat{y}_i)^2 &(25)\\
&= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i &(26)\\
\textbf{Total sum of squares} \quad SST &= \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2 &(27)\\
\textbf{Estimated variance} \quad \hat{\sigma}^2 &= \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = s^2 &(28)\\
\textbf{Regression sum of squares} \quad SSR &= SST - SSE &(29)
\end{aligned}$$

Summary of variance measures

$$
\begin{aligned}
\textbf{Error sum of squares} \quad SSE &= \sum (y_i - \hat{y}_i)^2 && (25)\\
&= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i && (26)\\
\textbf{Total sum of squares} \quad SST &= \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2 && (27)\\
\textbf{Estimated variance} \quad \hat{\sigma}^2 &= \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = s^2 && (28)\\
\textbf{Regression sum of squares} \quad SSR &= SST - SSE && (29)\\
\textbf{Coefficient of determination} \quad R^2 &= 1 - \frac{SSE}{SST} = \frac{SSR}{SST} && (30)
\end{aligned}
$$

Note

The variance estimate $\hat{\sigma}^2$ is also defined as the *conditional variance*, $\mathbb{V}(Y|X = x)$