Definitions and assumptions

Tiago M. Magalhães

Department of Statistics - ICE-UFJF

Juiz de Fora, August 28, 2026

Outline

Outline

Introduction

The first concepts of regression were proposed by Galton (1889), applied in Anthropometry, with the objective of establishing the dependency relationships between the height of the parents (**explanatory variable** or covariate)

Introduction

The first concepts of regression were proposed by Galton (1889), applied in Anthropometry, with the objective of establishing the dependency relationships between the height of the parents (**explanatory variable** or covariate) and the height of the children (variable of interest or **response variable** or outcome).

Introduction

The first concepts of regression were proposed by Galton (1889), applied in Anthropometry, with the objective of establishing the dependency relationships between the height of the parents (**explanatory variable** or covariate) and the height of the children (variable of interest or **response variable** or outcome).

Introduction

However, these concepts can be applied in any context, such as:

- the need for a company to analyze the factors that may affect sales;

Introduction

However, these concepts can be applied in any context, such as:

- the need for a company to analyze the factors that may affect sales;
- predicting the mortality of a person after contracting a disease, based on their characteristics, such as age and pre-existing diseases.

Introduction

However, these concepts can be applied in any context, such as:

- the need for a company to analyze the factors that may affect sales;
- predicting the mortality of a person after contracting a disease, based on their characteristics, such as age and pre-existing diseases.

Introduction

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. In fact, regression analysis may be the most widely used statistical technique (Montgomery et al., 2021).

Introduction

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. In fact, regression analysis may be the most widely used statistical technique (Montgomery et al., 2021).

Outline

Linear regression model

Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent random variables, with $Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2)$, such that

$$\mu_\ell = \mathbb{E}(Y_\ell) = \beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p},$$

Linear regression model

Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent random variables, with $Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2)$, such that

$$\mu_\ell = \mathbb{E}(Y_\ell) = \beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p},$$

where $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ are known variables (quantities), $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ are unknown parameters to be estimated, and σ^2 is the variance of Y_ℓ, $\ell = 1, 2, \ldots, n$, also unknown and to be estimated.

Linear regression model

Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent random variables, with $Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2)$, such that

$$\mu_\ell = \mathbb{E}(Y_\ell) = \beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p},$$

where $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ are known variables (quantities), $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ are unknown parameters to be estimated, and σ^2 is the variance of Y_ℓ, $\ell = 1, 2, \ldots, n$, also unknown and to be estimated.

Linear regression model

We can assume that

$$Y_\ell = \beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p} + \varepsilon_\ell, \tag{1}$$

$$= \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,$$

where $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$, and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with mean zero and common variance σ^2.

Linear regression model

We can assume that

$$Y_\ell = \beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p} + \varepsilon_\ell, \qquad (1)$$

$$= \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,$$

where $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$, and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with mean zero and common variance σ^2.

Remark: this formulation is always possible for location-scale families.

Linear regression model

We can assume that

$$Y_\ell = \beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p} + \varepsilon_\ell, \tag{1}$$
$$= \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,$$

where $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$, and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with mean zero and common variance σ^2.

Remark: this formulation is always possible for location-scale families.

Linear regression model

Equation (1) is what we define as the **linear regression model** (LRM).

Remark: Many authors are not directly concerned with the distribution of Y_ℓ, but rather with that of ε_ℓ, $\ell = 1, 2, \ldots, n$.

Linear regression model

Equation (1) is what we define as the **linear regression model** (LRM).

Remark: Many authors are not directly concerned with the distribution of Y_ℓ, but rather with that of ε_ℓ, $\ell = 1, 2, \ldots, n$.

Interpretation of the model parameters

In the LRM, the vectors $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ and $\boldsymbol{x}_\ell$, $\ell = 1, 2, \ldots, n$, are called the response vector and predictor vector, respectively, and the parameters belonging to the vector β are called regression coefficients.

Interpretation of the model parameters

In the LRM, the vectors $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ and $\boldsymbol{x}_\ell$, $\ell = 1, 2, \ldots, n$, are called the response vector and predictor vector, respectively, and the parameters belonging to the vector $\boldsymbol{\beta}$ are called regression coefficients.

Interpretation of the model parameters

Parameter β_m, $m = 1, 2, \ldots, p$

It is the expected increase in the response variable when we increase x_m by one unit while keeping the remaining predictor variables constant.

Interpretation of the model parameters

Parameter β_m, $m = 1, 2, \ldots, p$

It is the expected increase in the response variable when we increase x_m by one unit while keeping the remaining predictor variables constant.

Interpretation of the model parameters

Intercept

It is the expected value of the response variable when all $p-1$ predictor variables are equal to 0.

Interpretation of the model parameters

Intercept

It is the expected value of the response variable when all $p - 1$ predictor variables are equal to 0.

Remark: For (1) to be an LRM with an intercept, it is enough to assume that $x_{\ell 1} = 1$, $\forall \ell = 1, 2, \ldots, n$.

Interpretation of the model parameters

Intercept

It is the expected value of the response variable when all $p-1$ predictor variables are equal to 0.

Remark: For (1) to be an LRM with an intercept, it is enough to assume that $x_{\ell 1} = 1$, $\forall \ell = 1, 2, \ldots, n$.

Simple linear regression model

When an LRM can be written as follows:

Simple linear regression model

When an LRM can be written as follows:

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell, \qquad (2)$$

$$= \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,$$

Simple linear regression model

When an LRM can be written as follows:

$$
\begin{aligned}
Y_\ell &= \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell, \\
&= \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,
\end{aligned}
\tag{2}
$$

where $\boldsymbol{x}_\ell = (1, x_{\ell 2})^\top$, $\ell = 1, 2, \ldots, n$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$.

Simple linear regression model

When an LRM can be written as follows:

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell, \qquad (2)$$
$$= \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,$$

where $\boldsymbol{x}_\ell = (1, x_{\ell 2})^\top$, $\ell = 1, 2, \ldots, n$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$. We call (2) the **simple linear regression model** (SLM).

Simple linear regression model

When an LRM can be written as follows:

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell, \qquad (2)$$
$$= \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,$$

where $\boldsymbol{x}_\ell = (1, x_{\ell 2})^\top$, $\ell = 1, 2, \ldots, n$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$. We call (2) the **simple linear regression model** (SLM).

Matrix form

An LRM can be written in matrix form,

Matrix form

An LRM can be written in matrix form, as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon, \qquad (3)$$

Matrix form

An LRM can be written in matrix form, as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, where $\mathbf{0}$ is the zero vector of dimension n, $\boldsymbol{I}_n$ is the identity matrix of order n, and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ or

Matrix form

An LRM can be written in matrix form, as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, where $\mathbf{0}$ is the zero vector of dimension n, $\boldsymbol{I}_n$ is the identity matrix of order n, and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ or

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

Matrix form

An LRM can be written in matrix form, as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, where $\mathbf{0}$ is the zero vector of dimension n, $\boldsymbol{I}_n$ is the identity matrix of order n, and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$

or

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

Matrix form

The matrix $\boldsymbol{X}$ has n rows (sample size) and p columns (number of predictor variables, covariates), of rank p ($n \geq p$), and $\boldsymbol{X}$ is called the design matrix.

The matrix form is quite useful for computational implementation, as it avoids looping.

Matrix form

The matrix $\boldsymbol{X}$ has n rows (sample size) and p columns (number of predictor variables, covariates), of rank p ($n \geq p$), and $\boldsymbol{X}$ is called the design matrix.

The matrix form is quite useful for computational implementation, as it avoids looping.

Normality

For classical finite-sample inferential procedures to be applied, we need to assume that $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, in (1) and (3), respectively. With this assumption, the LRM is called the **normal linear regression model** (NLM).

Normality

For classical finite-sample inferential procedures to be applied, we need to assume that $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, in (1) and (3), respectively. With this assumption, the LRM is called the **normal linear regression model** (NLM).

Note that, $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2) \Rightarrow Y_\ell \sim \mathcal{N}(\boldsymbol{x}_\ell^\top \beta, \sigma^2)$, $\ell = 1, 2, \ldots, n$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n) \Rightarrow \boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n)$.

Normality

For classical finite-sample inferential procedures to be applied, we need to assume that $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, in (1) and (3), respectively. With this assumption, the LRM is called the **normal linear regression model** (NLM).

Note that, $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2) \Rightarrow Y_\ell \sim \mathcal{N}(\boldsymbol{x}_\ell^\top \boldsymbol{\beta}, \sigma^2)$, $\ell = 1, 2, \ldots, n$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n) \Rightarrow \boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$.

Examples

Example 1. Westwood Company example (Neter et al., 1983, p. 36), also available in Table 1.

Examples

Example 1. Westwood Company example (Neter et al., 1983, p. 36), also available in Table 1.

Table 1: Data on lot size and number of man-hours.

Production run	1	2	3	4	5	6	7	8	9	10
Lot size	30	20	60	80	40	50	60	30	70	60
Man-hours	73	50	128	170	87	108	135	69	148	132

Examples

Example 1. Westwood Company example (Neter et al., 1983, p. 36), also available in Table 1.

Table 1: Data on lot size and number of man-hours.

Production run	1	2	3	4	5	6	7	8	9	10
Lot size	30	20	60	80	40	50	60	30	70	60
Man-hours	73	50	128	170	87	108	135	69	148	132

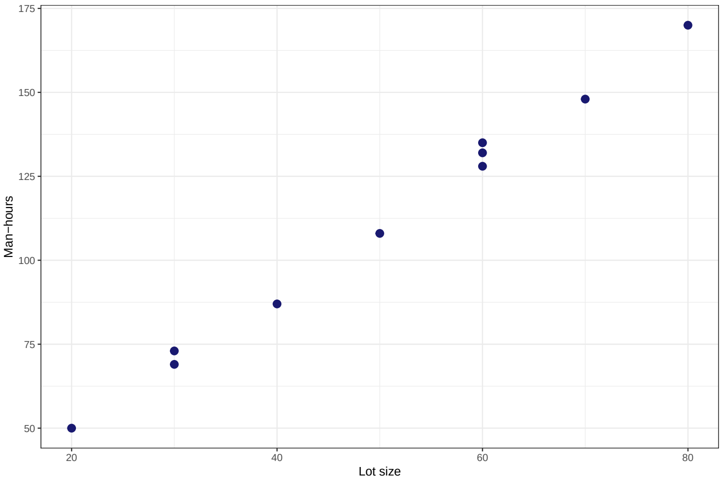

Figure 1: Scatterplot for the Westwood Company example.

Examples

From Figure 1, we could assume an LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell,$$

Examples

From Figure 1, we could assume an LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell,$$

where Y_ℓ: man-hours, $x_{\ell 2}$: lot size, $\ell = 1, 2, \ldots, 10$.

Examples

From Figure 1, we could assume an LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell,$$

where Y_ℓ: man-hours, $x_{\ell 2}$: lot size, $\ell = 1, 2, \ldots, 10$.

Examples

Example 2. The rocket propellant data (Montgomery et al., 2021, p. 15), also available in Table 2.

Examples

Example 2. The rocket propellant data (Montgomery et al., 2021, p. 15), also available in Table 2.

Table 2: Rocket propellant.

Shear strength (psi)	Age of propellant (weeks)	Shear strength (psi)	Age of propellant (weeks)
2158.70	15.50	2165.20	13.00
1678.15	23.75	2399.55	3.75
2316.00	8.00	1779.80	25.00
2061.30	17.00	2336.75	9.75
2207.50	5.50	1765.30	22.00
1708.30	19.00	2053.50	18.00
1784.70	24.00	2414.40	6.00
2575.00	2.50	2200.50	12.50
2357.90	7.50	2654.20	2.00
2256.70	11.00	1753.70	21.50

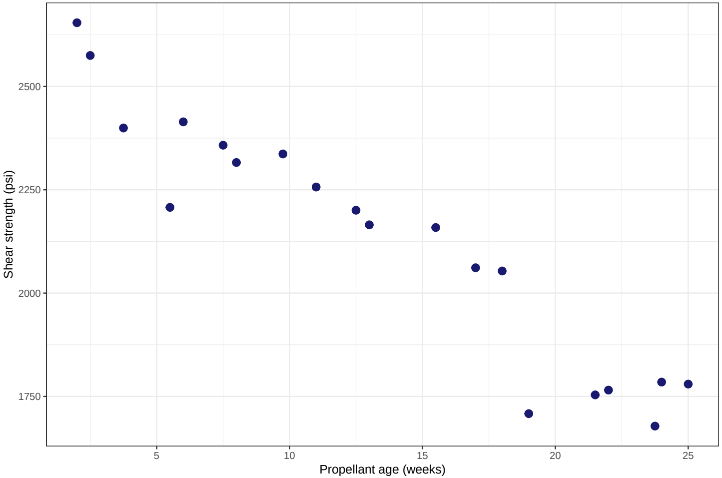

Figure 2: Scatterplot for the rocket propellant data.

Examples

From Figure 2, we could assume an LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell,$$

Examples

From Figure 2, we could assume an LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell,$$

where Y_ℓ: shear strength (psi), $x_{\ell 2}$: propellant age (weeks), $\ell = 1, 2, \ldots, 20$.

Examples

From Figure 2, we could assume an LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell,$$

where Y_ℓ: shear strength (psi), $x_{\ell 2}$: propellant age (weeks), $\ell = 1, 2, \ldots, 20$.

Outline

Standardization

In some situations, it is convenient to change the scale of the variables. For example, when:

- there is high correlation between predictor variables;

Standardization

In some situations, it is convenient to change the scale of the variables. For example, when:

- there is high correlation between predictor variables;
- the ranges of variation of the covariates are very different.

Standardization

In some situations, it is convenient to change the scale of the variables. For example, when:

- there is high correlation between predictor variables;
- the ranges of variation of the covariates are very different.

Standardization

The first possibility is as follows:

$$x^*_{\ell m} = x_{\ell m} - x^0_m,$$

Standardization

The first possibility is as follows:

$$x^*_{\ell m} = x_{\ell m} - x^0_m,$$

where $x^0_{\ell m}$ is a reference value, $\ell = 1, 2, \ldots, n$ and $m = 1, 2, \ldots, p$.

Standardization

The first possibility is as follows:

$$x^*_{\ell m} = x_{\ell m} - x^0_m,$$

where $x^0_{\ell m}$ is a reference value, $\ell = 1, 2, \ldots, n$ and $m = 1, 2, \ldots, p$.

Commonly, $x^0_m = \bar{x}_m$ is used, the sample mean of the *m*th covariate.

Standardization

The first possibility is as follows:

$$x_{\ell m}^{*} = x_{\ell m} - x_{m}^{0},$$

where $x_{\ell m}^{0}$ is a reference value, $\ell = 1, 2, \ldots, n$ and $m = 1, 2, \ldots, p$.

Commonly, $x_{m}^{0} = \bar{x}_{m}$ is used, the sample mean of the mth covariate.

Unit normal scaling

In the second possibility, called **unit normal scaling**, the procedure is as follows:

$$Y_\ell^* = \frac{Y_\ell - \bar{Y}}{S_Y} \text{ and } x_{\ell m}^* = \frac{x_{\ell m} - \bar{x}_m}{s_m},$$

Unit normal scaling

In the second possibility, called **unit normal scaling**, the procedure is as follows:

$$Y_\ell^* = \frac{Y_\ell - \bar{Y}}{S_Y} \text{ and } x_{\ell m}^* = \frac{x_{\ell m} - \bar{x}_m}{s_m},$$

where S_Y and s_m are, respectively, the square roots of

$$S_Y^2 = \frac{\sum_{\ell=1}^{n}(Y_\ell - \bar{Y})^2}{n-1} \text{ and } s_m^2 = \frac{\sum_{\ell=1}^{n}(x_{\ell m} - \bar{x}_m)^2}{n-1},$$

$\ell = 1, 2, \ldots, n$ and $m = 1, 2, \ldots, p$.

Unit normal scaling

In the second possibility, called **unit normal scaling**, the procedure is as follows:

$$Y_\ell^* = \frac{Y_\ell - \bar{Y}}{S_Y} \text{ and } x_{\ell m}^* = \frac{x_{\ell m} - \bar{x}_m}{s_m},$$

where S_Y and s_m are, respectively, the square roots of

$$S_Y^2 = \frac{\sum_{\ell=1}^{n}(Y_\ell - \bar{Y})^2}{n-1} \text{ and } s_m^2 = \frac{\sum_{\ell=1}^{n}(x_{\ell m} - \bar{x}_m)^2}{n-1},$$

$\ell = 1, 2, \ldots, n$ and $m = 1, 2, \ldots, p$.

Unit normal scaling

Remark

All transformed regressors and transformed responses have sample mean zero and sample variance one.

Unit normal scaling

Remark

All transformed regressors and transformed responses have sample mean zero and sample variance one.

Unit length scaling

In the third possibility, called **unit length scaling**, the procedure is as follows:

$$Y_\ell^* = \frac{Y_\ell - \bar{Y}}{S_Y} \text{ and } x_{\ell m}^* = \frac{x_{\ell m} - \bar{x}_m}{s_m},$$

Unit length scaling

In the third possibility, called **unit length scaling**, the procedure is as follows:

$$Y_\ell^* = \frac{Y_\ell - \bar{Y}}{S_Y} \text{ and } x_{\ell m}^* = \frac{x_{\ell m} - \bar{x}_m}{s_m},$$

where S_Y and s_m are, respectively, the square roots of

$$S_Y^2 = \sum_{\ell=1}^{n} (Y_\ell - \bar{Y})^2 \text{ and } S_m^2 = \sum_{\ell=1}^{n} (x_{\ell m} - \bar{x}_m)^2,$$

$\ell = 1, 2, \ldots, n$ and $m = 1, 2, \ldots, p$.

Unit length scaling

In the third possibility, called **unit length scaling**, the procedure is as follows:

$$Y_\ell^* = \frac{Y_\ell - \bar{Y}}{S_Y} \text{ and } x_{\ell m}^* = \frac{x_{\ell m} - \bar{x}_m}{s_m},$$

where S_Y and s_m are, respectively, the square roots of

$$S_Y^2 = \sum_{\ell=1}^{n}(Y_\ell - \bar{Y})^2 \text{ and } S_m^2 = \sum_{\ell=1}^{n}(x_{\ell m} - \bar{x}_m)^2,$$

$\ell = 1, 2, \ldots, n$ and $m = 1, 2, \ldots, p$.

Unit length scaling

Remark

Each regressor and the response have sample mean zero and unit Euclidean norm,

Unit length scaling

Remark

Each regressor and the response have sample mean zero and unit Euclidean norm, that is,

$$\sqrt{\sum_{\ell=1}^{n}(x_{\ell m}^{*} - \bar{x}_{m}^{*})^2} = 1.$$

Unit length scaling

Remark

Each regressor and the response have sample mean zero and unit Euclidean norm, that is,

$$\sqrt{\sum_{\ell=1}^{n}(x_{\ell m}^{*}-\bar{x}_{m}^{*})^{2}}=1.$$

Standardization

Final remark

Note that, in a model with an intercept, the variable standardization procedure removes the intercept from the model.

Standardization

Final remark

Note that, in a model with an intercept, the variable standardization procedure removes the intercept from the model.

Outline

Qualitative covariates

It is not uncommon for qualitative (categorical) predictor variables to be used to explain the behavior of a response variable.

However, their use requires some additional care in their treatment.

Qualitative covariates

It is not uncommon for qualitative (categorical) predictor variables to be used to explain the behavior of a response variable.

However, their use requires some additional care in their treatment.

Qualitative covariates

Situation 1. Suppose the following LRM,

Qualitative covariates

Situation 1. Suppose the following LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell,$$

Qualitative covariates

Situation 1. Suppose the following LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell,$$

where $x_{\ell 2}$ is any continuous variable and $x_{\ell 3}$ is the sex of individual ℓ, $x_{\ell 3} \in \{\text{female}, \text{male}\}$, $\ell = 1, 2, \ldots, n$.

Qualitative covariates

Situation 1. Suppose the following LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell,$$

where $x_{\ell 2}$ is any continuous variable and $x_{\ell 3}$ is the sex of individual ℓ, $x_{\ell 3} \in \{\text{female}, \text{male}\}$, $\ell = 1, 2, \ldots, n$.

Qualitative covariates

Suppose also that

$$x_{\ell 3} = \begin{cases} 1, & \text{if female} \\ 0, & \text{if male} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

Qualitative covariates

Suppose also that

$$x_{\ell 3} = \begin{cases} 1, & \text{if female} \\ 0, & \text{if male} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

The variable $x_{\ell 3}$, written in the form above, is called an indicator variable or **dummy variable**.

Qualitative covariates

Suppose also that

$$x_{\ell 3} = \begin{cases} 1, & \text{if female} \\ 0, & \text{if male} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

The variable $x_{\ell 3}$, written in the form above, is called an indicator variable or **dummy variable**. Now, note that,

$$x_{\ell 3} = \begin{cases} 1: & Y_\ell = (\beta_1 + \beta_3) + \beta_2 x_{\ell 2} + \varepsilon_\ell \\ 0: & Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell \end{cases}, \ \ell = 1, 2, \ldots, n.$$

Qualitative covariates

Suppose also that

$$x_{\ell 3} = \begin{cases} 1, & \text{if female} \\ 0, & \text{if male} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

The variable $x_{\ell 3}$, written in the form above, is called an indicator variable or **dummy variable**. Now, note that,

$$x_{\ell 3} = \begin{cases} 1: & Y_\ell = (\beta_1 + \beta_3) + \beta_2 x_{\ell 2} + \varepsilon_\ell \\ 0: & Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell \end{cases}, \ \ell = 1, 2, \ldots, n.$$

That is, the LRM has been converted into two SLMs.

Qualitative covariates

Suppose also that

$$x_{\ell 3} = \begin{cases} 1, & \text{if female} \\ 0, & \text{if male} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

The variable $x_{\ell 3}$, written in the form above, is called an indicator variable or **dummy variable**. Now, note that,

$$x_{\ell 3} = \begin{cases} 1: & Y_\ell = (\beta_1 + \beta_3) + \beta_2 x_{\ell 2} + \varepsilon_\ell \\ 0: & Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell \end{cases}, \ \ell = 1, 2, \ldots, n.$$

That is, the LRM has been converted into two SLMs.

Situation 2. Now, suppose the following LRM,

Qualitative covariates

Situation 2. Now, suppose the following LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell,$$

Qualitative covariates

Situation 2. Now, suppose the following LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell,$$

where $x_{\ell 2}$ is any continuous variable and $x_{\ell 3}$ is the educational level of individual ℓ, $x_{\ell 3} \in \{\text{elementary}, \text{secondary}, \text{higher education}\}$, $\ell = 1, 2, \ldots, n$.

Qualitative covariates

Situation 2. Now, suppose the following LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell,$$

where $x_{\ell 2}$ is any continuous variable and $x_{\ell 3}$ is the educational level of individual ℓ, $x_{\ell 3} \in \{\text{elementary}, \text{secondary}, \text{higher education}\}$, $\ell = 1, 2, \ldots, n$.

Qualitative covariates

For this situation, it is necessary to create two dummy variables:

$$x_{\ell 3} = \begin{cases} 1, & \text{if has elementary education} \\ 0, & \text{otherwise} \end{cases},$$

$$x_{\ell 4} = \begin{cases} 1, & \text{if has secondary education} \\ 0, & \text{otherwise} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

Qualitative covariates

For this situation, it is necessary to create two dummy variables:

$$x_{\ell 3} = \begin{cases} 1, & \text{if has elementary education} \\ 0, & \text{otherwise} \end{cases},$$

$$x_{\ell 4} = \begin{cases} 1, & \text{if has secondary education} \\ 0, & \text{otherwise} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

Qualitative covariates

Now, note that,

$$\begin{cases} \text{Elementary} & (x_{\ell 3} = 1, x_{\ell 4} = 0): \quad Y_\ell = (\beta_1 + \beta_3) + \beta_2 x_{\ell 2} + \varepsilon_\ell \\ \text{Secondary} & (x_{\ell 3} = 0, x_{\ell 4} = 1): \quad Y_\ell = (\beta_1 + \beta_4) + \beta_2 x_{\ell 2} + \varepsilon_\ell \\ \text{Higher education} & (x_{\ell 3} = 0, x_{\ell 4} = 0): \quad Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell \end{cases},$$

$\ell = 1, 2, \ldots, n.$

Qualitative covariates

Now, note that,

$$\begin{cases} \text{Elementary} & (x_{\ell 3}=1, x_{\ell 4}=0): \quad Y_\ell = (\beta_1+\beta_3)+\beta_2 x_{\ell 2}+\varepsilon_\ell \\ \text{Secondary} & (x_{\ell 3}=0, x_{\ell 4}=1): \quad Y_\ell = (\beta_1+\beta_4)+\beta_2 x_{\ell 2}+\varepsilon_\ell \\ \text{Higher education} & (x_{\ell 3}=0, x_{\ell 4}=0): \quad Y_\ell = \beta_1+\beta_2 x_{\ell 2}+\varepsilon_\ell \end{cases},$$

$\ell = 1, 2, \ldots, n$. That is, the LRM has been converted into three SLMs. Most importantly, the inclusion of a qualitative variable changes the design matrix.

Qualitative covariates

Now, note that,

$$\begin{cases} \text{Elementary} & (x_{\ell 3} = 1, x_{\ell 4} = 0): \quad Y_\ell = (\beta_1 + \beta_3) + \beta_2 x_{\ell 2} + \varepsilon_\ell \\ \text{Secondary} & (x_{\ell 3} = 0, x_{\ell 4} = 1): \quad Y_\ell = (\beta_1 + \beta_4) + \beta_2 x_{\ell 2} + \varepsilon_\ell \\ \text{Higher education} & (x_{\ell 3} = 0, x_{\ell 4} = 0): \quad Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \varepsilon_\ell \end{cases},$$

$\ell = 1, 2, \ldots, n$. That is, the LRM has been converted into three SLMs. Most importantly, the inclusion of a qualitative variable changes the design matrix.

$$\boldsymbol{X}_{\mathrm{Sit.1}} = \begin{bmatrix} 1 & x_{12} & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_{k2} & 0 \\ 1 & x_{(k+1)2} & 1 \\ \vdots & \vdots & \vdots \\ 1 & x_{n2} & 1 \end{bmatrix} \text{ and } \boldsymbol{X}_{\mathrm{Sit.2}} = \begin{bmatrix} 1 & x_{12} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{k_1 2} & 0 & 0 \\ 1 & x_{(k_1+1)2} & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{k_2 2} & 0 & 1 \\ 1 & x_{(k_2+1)2} & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & 1 & 0 \end{bmatrix}.$$

Qualitative covariates

If a predictor variable has k categories, $k - 1$ dummy variables will be required.

Examples

Example 3. Dataset from Walker (1962). The author studied the sounds of male crickets in the USA, measuring the number of pulses per second of 31 crickets at different temperatures, recorded in degrees Fahrenheit.

The data are reproduced in McDonald (2009), where the temperatures are reported in degrees Celsius; see Table 3.

Examples

Example 3. Dataset from Walker (1962). The author studied the sounds of male crickets in the USA, measuring the number of pulses per second of 31 crickets at different temperatures, recorded in degrees Fahrenheit.

The data are reproduced in McDonald (2009), where the temperatures are reported in degrees Celsius; see Table 3.

Table 3: Cricket pulse rate.

Spec.	Temp.	Pulses	Spec.	Temp.	Pulses	Spec.	Temp.	Pulses
ex	20.8	67.9	ex	29.0	100.8	niv	21.0	58.9
ex	20.8	65.1	ex	30.4	99.3	niv	22.1	60.7
ex	24.0	77.3	ex	30.4	101.7	niv	23.5	69.8
ex	24.0	78.7	niv	17.2	44.3	niv	24.2	70.9
ex	24.0	79.4	niv	18.3	47.2	niv	25.9	76.2
ex	24.0	80.4	niv	18.3	47.6	niv	26.5	76.1
ex	26.2	85.8	niv	18.3	49.6	niv	26.5	77.0
ex	26.2	86.6	niv	18.9	50.3	niv	26.5	77.7
ex	26.2	87.5	niv	18.9	51.8	niv	28.6	84.7
ex	26.2	89.1	niv	20.4	60.0			
ex	28.4	98.6	niv	21.0	58.5			

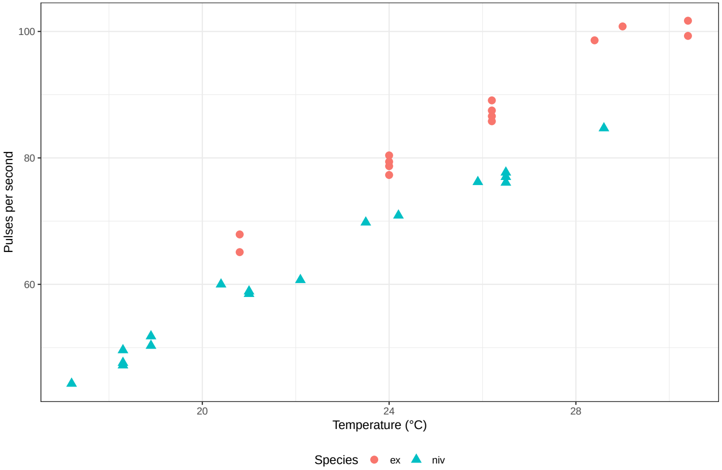

Figure 3: Scatterplot of the number of pulses per second versus temperature.

Examples

From Figure 3, we could assume an LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell,$$

Examples

From Figure 3, we could assume an LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell,$$

where Y_ℓ denotes the number of pulses per second, $x_{\ell 2}$ denotes temperature (°C), and $x_{\ell 3}$ denotes cricket species, with $x_{\ell 3} \in \{\textit{Oecanthus exclamationis (ex)}, \textit{Oecanthus niveus (niv)}\}$, $\ell = 1, 2, \ldots, 31$.

Examples

From Figure 3, we could assume an LRM,

$$Y_\ell = \beta_1 + \beta_2 x_{\ell 2} + \beta_3 x_{\ell 3} + \varepsilon_\ell,$$

where Y_ℓ denotes the number of pulses per second, $x_{\ell 2}$ denotes temperature ($^\circ$C), and $x_{\ell 3}$ denotes cricket species, with $x_{\ell 3} \in \{\textit{Oecanthus exclamationis (ex)}, \textit{Oecanthus niveus (niv)}\}$, $\ell = 1, 2, \ldots, 31$.

Outline

Simulation study

Suppose the following NLM,

$$Y_\ell = 1 + x_{\ell 2} + \varepsilon_\ell,$$

Simulation study

Suppose the following NLM,

$$Y_\ell = 1 + x_{\ell 2} + \varepsilon_\ell,$$

where $x_{\ell 2}$ is a known quantity and $\varepsilon_\ell \sim \mathcal{N}(0, 4)$, $\ell = 1, 2, \ldots, n$.

Simulation study

Suppose the following NLM,

$$Y_\ell = 1 + x_{\ell 2} + \varepsilon_\ell,$$

where $x_{\ell 2}$ is a known quantity and $\varepsilon_\ell \sim \mathcal{N}(0, 4)$, $\ell = 1, 2, \ldots, n$. Note that $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$ and we will assume that $x_{\ell 2} \sim$ uniform on the interval (0,1).

Simulation study

Suppose the following NLM,

$$Y_\ell = 1 + x_{\ell 2} + \varepsilon_\ell,$$

where $x_{\ell 2}$ is a known quantity and $\varepsilon_\ell \sim \mathcal{N}(0, 4)$, $\ell = 1, 2, \ldots, n$. Note that $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$ and we will assume that $x_{\ell 2} \sim$ uniform on the interval (0,1).

We will conduct a Monte Carlo study, with 5,000 replications, where we will evaluate $\text{Cov}(Y, x_2)$, for $n \in \{10, 20, 40, 80, 160\}$.

Simulation study

Suppose the following NLM,

$$Y_\ell = 1 + x_{\ell 2} + \varepsilon_\ell,$$

where $x_{\ell 2}$ is a known quantity and $\varepsilon_\ell \sim \mathcal{N}(0,4)$, $\ell = 1, 2, \ldots, n$. Note that $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$ and we will assume that $x_{\ell 2} \sim$ uniform on the interval (0,1).

We will conduct a Monte Carlo study, with 5,000 replications, where we will evaluate $\mathrm{Cov}(Y, x_2)$, for $n \in \{10, 20, 40, 80, 160\}$.

Simulation study

That is,

1. We fix the parameters, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ and $n = 10$;

Simulation study

That is,

1. We fix the parameters, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ and $n = 10$;
2. We generate 10 values of $\boldsymbol{x}_2$ from a uniform(0, 1);

Simulation study

That is,

1. We fix the parameters, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ and $n = 10$;
2. We generate 10 values of $\boldsymbol{x}_2$ from a uniform(0, 1);
3. For each replication, we generate 10 values of $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;

Simulation study

That is,

1. We fix the parameters, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ and $n = 10$;
2. We generate 10 values of $\boldsymbol{x}_2$ from a uniform(0, 1);
3. For each replication, we generate 10 values of $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;
4. Also for each replication, we calculate $\text{Cov}(Y, x_2)$;

Simulation study

That is,

1. We fix the parameters, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ and $n = 10$;
2. We generate 10 values of $\boldsymbol{x}_2$ from a uniform(0, 1);
3. For each replication, we generate 10 values of $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;
4. Also for each replication, we calculate $\text{Cov}(Y, x_2)$;
5. With the 5,000 replications, we calculate the mean of the 5,000 covariances.

Simulation study

That is,

1. We fix the parameters, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ and $n = 10$;
2. We generate 10 values of $\boldsymbol{x}_2$ from a uniform(0, 1);
3. For each replication, we generate 10 values of $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;
4. Also for each replication, we calculate $\text{Cov}(Y, x_2)$;
5. With the 5,000 replications, we calculate the mean of the 5,000 covariances.

The above experiment was repeated 4 more times, changing only the sample size to $20, 40, 80, 160$.

Simulation study

That is,

1. We fix the parameters, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ and $n = 10$;
2. We generate 10 values of $\boldsymbol{x}_2$ from a uniform(0, 1);
3. For each replication, we generate 10 values of $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;
4. Also for each replication, we calculate $\text{Cov}(Y, x_2)$;
5. With the 5,000 replications, we calculate the mean of the 5,000 covariances.

The above experiment was repeated 4 more times, changing only the sample size to $20, 40, 80, 160$.

Simulation study

Table 4: Results.

n	Covariance	SD
10	0.063	0.167
20	0.090	0.137
40	0.068	0.085
80	0.086	0.066
160	0.088	0.047

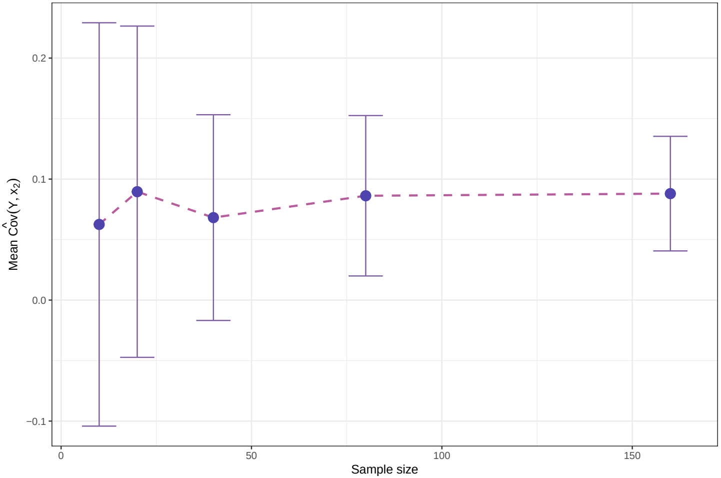

Figure 4: Mean covariance as a function of sample size.

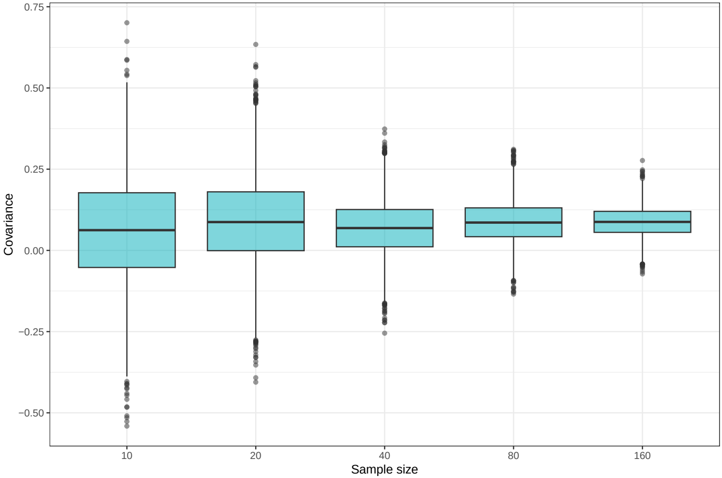

Figure 5: Boxplots of the simulated covariances for different sample sizes.

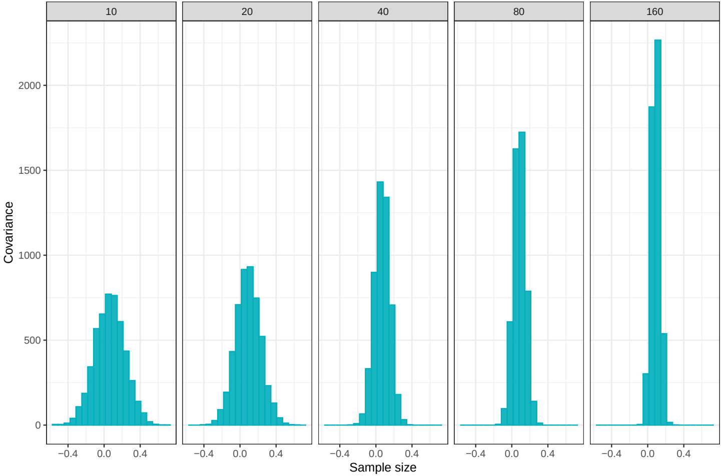

Figure 6: Histograms of the simulated covariances for different sample sizes.

Simulation study

Comments

From Figures 4 to 6, we can see that the estimates of the covariances between Y and x_2 seem to converge to a **certain value**.

Simulation study

Comments

From Figures 4 to 6, we can see that the estimates of the covariances between Y and x_2 seem to converge to a **certain value**. As expected, as the sample size increases, the variability of the covariance estimates decreases.

Simulation study

Comments

From Figures 4 to 6, we can see that the estimates of the covariances between Y and x_2 seem to converge to a **certain value**. As expected, as the sample size increases, the variability of the covariance estimates decreases.

Outline

Final remarks

For a better understanding of the concepts of linear regression models, see Chapter 1 of Kutner et al. (2005) and Montgomery et al. (2021).

Outline

References

Galton, F. (1889). *Natural Inheritance*. Macmillan, London.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied linear statistical models*. McGraw-Hill Irwin, Chicago, 5th edition.

McDonald, J. H. (2009). *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, Maryland, 2nd edition.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. Wiley, New York, 6th edition.

References

Neter, J., Wasserman, W., and Kutner, M. H. (1983). *Applied linear regression models*. Richard D. Irwin Inc, Homewood, Illinois.

Walker, T. J. (1962). The Taxonomy and Calling Songs of United States Tree Crickets (Orthoptera: Gryllidae: Oecanthinae). I. The Genus Neoxabea and the niveus and varicornis Groups of the Genus Oecanthus. *Annals of the Entomological Society of America*, 55(3):303–322.

Thank you!

✉ tiago.magalhaes@ufjf.br

ufjf.br/tiago_magalhaes

Department of Statistics, Room 319