Estimation

Tiago M. Magalhães

Department of Statistics - ICE-UFJF

Juiz de Fora, September 02, 2026

Outline

Outline

Linear regression model

Suppose that $Y_1, Y_2, \ldots, Y_n$ satisfy

$$Y_\ell = \boldsymbol{x}_\ell^\top \beta + \varepsilon_\ell, \tag{1}$$

Linear regression model

Suppose that $Y_1, Y_2, \ldots, Y_n$ satisfy

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \tag{1}$$

where $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ is known, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is a vector of unknown parameters to be estimated, $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with mean zero and common variance σ^2, also unknown, to be estimated.

Linear regression model

Suppose that $Y_1, Y_2, \ldots, Y_n$ satisfy

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \tag{1}$$

where $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ is known, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is a vector of unknown parameters to be estimated, $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with mean zero and common variance σ^2, also unknown, to be estimated.

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;
  - have constant variance;

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;
  - have constant variance;
  - are uncorrelated.

Assumptions

In summary, the **assumptions** of a linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;
  - have constant variance;
  - are uncorrelated.

Matrix form

Equation (1), which we defined as the **linear regression model** (LRM), can be written in matrix form as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon, \tag{2}$$

Matrix form

Equation (1), which we defined as the **linear regression model** (LRM), can be written in matrix form as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, where $\mathbf{0}$ is the zero vector of dimension n, $\boldsymbol{I}_n$ is the identity matrix of order n and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$

Matrix form

Equation (1), which we defined as the **linear regression model** (LRM), can be written in matrix form as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, where $\mathbf{0}$ is the zero vector of dimension n, $\boldsymbol{I}_n$ is the identity matrix of order n and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$

Matrix form

or

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

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

Matrix form

or

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

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

Normal linear regression model

To apply inferential procedures, 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 (2), respectively. The LRM is then referred to as the **normal linear regression model** (NLM).

Normal linear regression model

To apply inferential procedures, 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 (2), respectively. The LRM is then referred to as the **normal linear regression model** (NLM).

Outline

Least squares method

Definition

The least squares estimators (LSEs) provide estimates of the unknown parameters by **minimizing** the sum of squared errors.

Least squares method

Definition

The least squares estimators (LSEs) provide estimates of the unknown parameters by **minimizing** the sum of squared errors.

Least squares method

If $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)^\top$ is the LSE of $\boldsymbol{\beta}$, then

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}),$$

where

$$Q(\boldsymbol{\beta}) = \sum_{\ell=1}^{n} \left[Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p})\right]^2 . \tag{3}$$

Least squares method

Differentiating (3), with respect to the parameter β_m, $m = 1, 2, \ldots, p$, we have:

$$\frac{\partial Q(\beta)}{\partial \beta_m} = -2 \sum_{\ell=1}^{n} \left[Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p}) \right] x_{\ell m}. \qquad (4)$$

Least squares method

Differentiating (3), with respect to the parameter β_m, $m = 1, 2, \ldots, p$, we have:

$$\frac{\partial Q(\beta)}{\partial \beta_m} = -2 \sum_{\ell=1}^{n} \left[Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p}) \right] x_{\ell m}. \qquad (4)$$

Setting (4) equal to zero,

Least squares method

Differentiating (3), with respect to the parameter β_m, $m = 1, 2, \ldots, p$, we have:

$$\frac{\partial Q(\beta)}{\partial \beta_m} = -2 \sum_{\ell=1}^{n} \left[Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p}) \right] x_{\ell m}. \qquad (4)$$

Setting (4) equal to zero, we have

$$\hat{\beta}_1 \sum_{\ell=1}^{n} x_{\ell 1} x_{\ell m} + \cdots + \hat{\beta}_m \sum_{\ell=1}^{n} x_{\ell m}^2 + \cdots + \hat{\beta}_p \sum_{\ell=1}^{n} x_{\ell p} x_{\ell m} = \sum_{\ell=1}^{n} x_{\ell m} Y_\ell. \qquad (5)$$

Least squares method

Differentiating (3), with respect to the parameter β_m, $m = 1, 2, \ldots, p$, we have:

$$\frac{\partial Q(\beta)}{\partial \beta_m} = -2 \sum_{\ell=1}^{n} \left[Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p}) \right] x_{\ell m}. \quad (4)$$

Setting (4) equal to zero, we have

$$\hat{\beta}_1 \sum_{\ell=1}^{n} x_{\ell 1} x_{\ell m} + \cdots + \hat{\beta}_m \sum_{\ell=1}^{n} x_{\ell m}^2 + \cdots + \hat{\beta}_p \sum_{\ell=1}^{n} x_{\ell p} x_{\ell m} = \sum_{\ell=1}^{n} x_{\ell m} Y_\ell. \quad (5)$$

Least squares method

The p equations in (5) are called the **normal equations**. The **least squares estimator** is therefore given by:

Least squares method

The p equations in (5) are called the **normal equations**. The **least squares estimator** is therefore given by:

$$\widehat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}. \tag{6}$$

Least squares method

The p equations in (5) are called the **normal equations**. The **least squares estimator** is therefore given by:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}. \tag{6}$$

Least squares method

An alternative way to obtain (6) is to observe that (3) can be written as:

$$Q(\beta) = (\boldsymbol{Y} - \boldsymbol{X}\beta)^\top (\boldsymbol{Y} - \boldsymbol{X}\beta).$$

Least squares method

An alternative way to obtain (6) is to observe that (3) can be written as:

$$Q(\beta) = (\boldsymbol{Y} - \boldsymbol{X}\beta)^\top (\boldsymbol{Y} - \boldsymbol{X}\beta).$$

and differentiate $Q(\beta)$ with respect to β.

Least squares method

An alternative way to obtain (6) is to observe that (3) can be written as:

$$Q(\beta) = (\boldsymbol{Y} - \boldsymbol{X}\beta)^\top (\boldsymbol{Y} - \boldsymbol{X}\beta).$$

and differentiate $Q(\beta)$ with respect to β.

Least squares method

First, we can observe:

$$\begin{aligned} Q(\beta) &= (\boldsymbol{Y}^\top - \beta^\top \boldsymbol{X}^\top)(\boldsymbol{Y} - \boldsymbol{X}\beta) \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{Y}^\top \boldsymbol{X}\beta - \beta^\top \boldsymbol{X}^\top \boldsymbol{Y} + \beta^\top \boldsymbol{X}^\top \boldsymbol{X}\beta \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - 2\beta^\top \boldsymbol{X}^\top \boldsymbol{Y} + \beta^\top \boldsymbol{X}^\top \boldsymbol{X}\beta. \end{aligned}$$

Now, differentiating Q with respect to β, we have:

$$\frac{\partial Q(\beta)}{\partial \beta} = -2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\beta. \tag{7}$$

Least squares method

First, we can observe:

$$\begin{aligned} Q(\beta) &= (\boldsymbol{Y}^\top - \beta^\top \boldsymbol{X}^\top)(\boldsymbol{Y} - \boldsymbol{X}\beta) \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{Y}^\top \boldsymbol{X}\beta - \beta^\top \boldsymbol{X}^\top \boldsymbol{Y} + \beta^\top \boldsymbol{X}^\top \boldsymbol{X}\beta \\ &= \boldsymbol{Y}^\top \boldsymbol{Y} - 2\beta^\top \boldsymbol{X}^\top \boldsymbol{Y} + \beta^\top \boldsymbol{X}^\top \boldsymbol{X}\beta. \end{aligned}$$

Now, differentiating Q with respect to β, we have:

$$\frac{\partial Q(\beta)}{\partial \beta} = -2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\beta. \tag{7}$$

Least squares method

Setting the derivative equal to the zero vector, we have:

$$
\begin{aligned}
-2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\widehat{\beta} &= \mathbf{0} \\
\Rightarrow \boldsymbol{X}^\top \boldsymbol{X}\widehat{\beta} &= \boldsymbol{X}^\top \boldsymbol{Y} \\
\Rightarrow \widehat{\beta} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}.
\end{aligned}
$$

Least squares method

Remark

The LSE exists if the inverse $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists.

Least squares method

Remark

The LSE exists if the inverse $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists. The matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists if the columns of $\boldsymbol{X}$ are linearly independent,

Least squares method

Remark

The LSE exists if the inverse $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists. The matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists if the columns of $\boldsymbol{X}$ are linearly independent, that is, if no column of $\boldsymbol{X}$ can be expressed as a linear combination of the others.

Least squares method

Remark

The LSE exists if the inverse $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists. The matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists if the columns of $\boldsymbol{X}$ are linearly independent, that is, if no column of $\boldsymbol{X}$ can be expressed as a linear combination of the others.

Computational cost of the LSE

The main computational costs of solving the normal equations associated with (6) are:

Computational cost of the LSE

The main computational costs of solving the normal equations associated with (6) are:

- forming $\boldsymbol{X}^\top \boldsymbol{X}$ requires $\mathcal{O}(np^2)$ operations;

Computational cost of the LSE

The main computational costs of solving the normal equations associated with (6) are:

- forming $\boldsymbol{X}^\top \boldsymbol{X}$ requires $\mathcal{O}(np^2)$ operations;
- solving the resulting $p \times p$ system requires $\mathcal{O}(p^3)$ operations.

Computational cost of the LSE

The main computational costs of solving the normal equations associated with (6) are:

- forming $\boldsymbol{X}^\top \boldsymbol{X}$ requires $\mathcal{O}(np^2)$ operations;
- solving the resulting $p \times p$ system requires $\mathcal{O}(p^3)$ operations.

Therefore, the overall computational cost is

$$\mathcal{O}(np^2 + p^3).$$

Computational cost of the LSE

The main computational costs of solving the normal equations associated with (6) are:

- forming $\boldsymbol{X}^\top \boldsymbol{X}$ requires $\mathcal{O}(np^2)$ operations;
- solving the resulting $p \times p$ system requires $\mathcal{O}(p^3)$ operations.

Therefore, the overall computational cost is

$$\mathcal{O}(np^2 + p^3).$$

When $n \gg p$, the dominant term is $\mathcal{O}(np^2)$.

Computational cost of the LSE

The main computational costs of solving the normal equations associated with (6) are:

- forming $\boldsymbol{X}^\top \boldsymbol{X}$ requires $\mathcal{O}(np^2)$ operations;
- solving the resulting $p \times p$ system requires $\mathcal{O}(p^3)$ operations.

Therefore, the overall computational cost is

$$\mathcal{O}(np^2 + p^3).$$

When $n \gg p$, the dominant term is $\mathcal{O}(np^2)$.

Properties of the LSE

The expectation of $\widehat{\beta}$ is given by

$$\begin{aligned}\mathbb{E}(\widehat{\beta}) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}(\boldsymbol{X}\beta + \varepsilon) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \left[\boldsymbol{X}\beta + \mathbb{E}(\varepsilon)\right] \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\beta = \beta.\end{aligned}$$

Properties of the LSE

The expectation of $\widehat{\beta}$ is given by

$$\begin{aligned}\mathbb{E}(\widehat{\beta}) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}(\boldsymbol{X}\beta + \varepsilon) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \left[\boldsymbol{X}\beta + \mathbb{E}(\varepsilon)\right] \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\beta = \beta.\end{aligned}$$

Properties of the LSE

And the variance of $\widehat{\beta}$ is given by

$$
\begin{aligned}
\text{Var}(\widehat{\beta}) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \text{Var}(\boldsymbol{Y}) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \text{Var}(\boldsymbol{X}\beta + \varepsilon) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \text{Var}(\varepsilon) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \sigma^2 \boldsymbol{I}_n \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.
\end{aligned}
$$

Properties of the LSE

And the variance of $\widehat{\beta}$ is given by

$$\begin{aligned}
\text{Var}(\widehat{\beta}) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \text{Var}(\boldsymbol{Y}) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \text{Var}(\boldsymbol{X}\beta + \varepsilon) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \text{Var}(\varepsilon) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \sigma^2 \boldsymbol{I}_n \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.
\end{aligned}$$

Properties of the LSE

By the Gauss-Markov Theorem (Plackett, 1949), the LSEs

- are unbiased;

Properties of the LSE

By the Gauss-Markov Theorem (Plackett, 1949), the LSEs

- are unbiased;
- and have the smallest variance among all linear unbiased estimators;

Properties of the LSE

By the Gauss-Markov Theorem (Plackett, 1949), the LSEs

- are unbiased;
- and have the smallest variance among all linear unbiased estimators;
- are the **best linear unbiased estimators (BLUEs)**.

Properties of the LSE

By the Gauss-Markov Theorem (Plackett, 1949), the LSEs

- are unbiased;
- and have the smallest variance among all linear unbiased estimators;
- are the **best linear unbiased estimators (BLUEs)**.

Fitted values

Thus, the fitted value of the ℓth observation is given by

Fitted values

Thus, the fitted value of the ℓth observation is given by

$$\hat{Y}_\ell = \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}},$$

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

Fitted values

Thus, the fitted value of the ℓth observation is given by

$$\hat{Y}_\ell = \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}},$$

$\ell = 1, 2, \ldots, n$. In matrix form,

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}.$$

Fitted values

Thus, the fitted value of the ℓth observation is given by

$$\hat{Y}_\ell = \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}},$$

$\ell = 1, 2, \ldots, n$. In matrix form,

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}.$$

The $n \times n$ matrix $\boldsymbol{H}$ is called the **hat matrix**.

Fitted values

Thus, the fitted value of the ℓth observation is given by

$$\hat{Y}_\ell = \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}},$$

$\ell = 1, 2, \ldots, n$. In matrix form,

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}.$$

The $n \times n$ matrix $\boldsymbol{H}$ is called the **hat matrix**.

Hat matrix

Properties of the matrix $\boldsymbol{H}$.

- $\boldsymbol{H}$ is symmetric and idempotent;

Hat matrix

Properties of the matrix $\boldsymbol{H}$.

- $\boldsymbol{H}$ is symmetric and idempotent;
- $(\boldsymbol{I}_n - \boldsymbol{H})$ is symmetric and idempotent;

Hat matrix

Properties of the matrix $\boldsymbol{H}$.

- $\boldsymbol{H}$ is symmetric and idempotent;
- $(\boldsymbol{I}_n - \boldsymbol{H})$ is symmetric and idempotent;
- $(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}$.

Hat matrix

Properties of the matrix $\boldsymbol{H}$.

- $\boldsymbol{H}$ is symmetric and idempotent;
- $(\boldsymbol{I}_n - \boldsymbol{H})$ is symmetric and idempotent;
- $(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}$.

Remark: a matrix $\boldsymbol{M}$ is said to be idempotent if and only if $\boldsymbol{MM} = \boldsymbol{M}$.

Hat matrix

Properties of the matrix $\boldsymbol{H}$.

- $\boldsymbol{H}$ is symmetric and idempotent;
- $(\boldsymbol{I}_n - \boldsymbol{H})$ is symmetric and idempotent;
- $(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}$.

Remark: a matrix $\boldsymbol{M}$ is said to be idempotent if and only if $\boldsymbol{MM} = \boldsymbol{M}$.

Residuals

It is the difference between the observed value and the corresponding fitted value.

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta} \\ &= \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}, \end{aligned}$$

Residuals

It is the difference between the observed value and the corresponding fitted value.

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}, \end{aligned}$$

where $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^\top$ and $e_\ell = Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}}$ is the ℓth residual, $\ell = 1, 2, \ldots, n$.

Residuals

It is the difference between the observed value and the corresponding fitted value.

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}, \end{aligned}$$

where $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^\top$ and $e_\ell = Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}}$ is the ℓth residual, $\ell = 1, 2, \ldots, n$.

Residuals

The expectation of $\boldsymbol{e}$ is given by

$$\begin{aligned}
\mathbb{E}(\boldsymbol{e}) &= (\boldsymbol{I}_n - \boldsymbol{H})\mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{I}_n - \boldsymbol{H})\mathbb{E}(\boldsymbol{X}\beta + \varepsilon) \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\left[\boldsymbol{X}\beta + \mathbb{E}(\varepsilon)\right] = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X}\beta \\
&= \boldsymbol{X}\beta - \boldsymbol{H}\boldsymbol{X}\beta \\
&= \boldsymbol{X}\beta - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\beta \\
&= \boldsymbol{X}\beta - \boldsymbol{X}\beta = \boldsymbol{0}.
\end{aligned}$$

Residuals

The expectation of $\boldsymbol{e}$ is given by

$$\begin{aligned}
\mathbb{E}(\boldsymbol{e}) &= (\boldsymbol{I}_n - \boldsymbol{H})\mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{I}_n - \boldsymbol{H})\mathbb{E}(\boldsymbol{X}\beta + \varepsilon) \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\left[\boldsymbol{X}\beta + \mathbb{E}(\varepsilon)\right] = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X}\beta \\
&= \boldsymbol{X}\beta - \boldsymbol{H}\boldsymbol{X}\beta \\
&= \boldsymbol{X}\beta - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\beta \\
&= \boldsymbol{X}\beta - \boldsymbol{X}\beta = \boldsymbol{0}.
\end{aligned}$$

Residuals

And the variance of $\boldsymbol{e}$ is given by

$$
\begin{aligned}
\text{Var}(\boldsymbol{e}) &= (\boldsymbol{I}_n - \boldsymbol{H})\text{Var}(\boldsymbol{Y})(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\text{Var}(\boldsymbol{X}\beta + \varepsilon)(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\text{Var}(\varepsilon)(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\sigma^2\boldsymbol{I}_n(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}).
\end{aligned}
$$

Residuals

And the variance of $\boldsymbol{e}$ is given by

$$\begin{aligned}
\text{Var}(\boldsymbol{e}) &= (\boldsymbol{I}_n - \boldsymbol{H})\text{Var}(\boldsymbol{Y})(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\text{Var}(\boldsymbol{X}\beta + \varepsilon)(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\text{Var}(\varepsilon)(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\sigma^2\boldsymbol{I}_n(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}).
\end{aligned}$$

Recall that $(\boldsymbol{I}_n - \boldsymbol{H})$ is idempotent and symmetric.

Residuals

And the variance of $\boldsymbol{e}$ is given by

$$\begin{aligned}
\text{Var}(\boldsymbol{e}) &= (\boldsymbol{I}_n - \boldsymbol{H})\text{Var}(\boldsymbol{Y})(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\text{Var}(\boldsymbol{X}\beta + \varepsilon)(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\text{Var}(\varepsilon)(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= (\boldsymbol{I}_n - \boldsymbol{H})\sigma^2\boldsymbol{I}_n(\boldsymbol{I}_n - \boldsymbol{H})^\top \\
&= \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}).
\end{aligned}$$

Recall that $(\boldsymbol{I}_n - \boldsymbol{H})$ is idempotent and symmetric.

Residuals

Let h_{ij} be the (i, j)th element of the matrix $\boldsymbol{H}$, $i, j = 1, 2, \ldots, n$, then we have:

- $\mathbb{E}(e_i) = 0$;

Residuals

Let h_{ij} be the (i,j)th element of the matrix $\boldsymbol{H}$, $i,j = 1,2,\ldots,n$, then we have:

- $\mathbb{E}(e_i) = 0$;
- $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$;

Residuals

Let h_{ij} be the (i,j)th element of the matrix $\boldsymbol{H}$, $i,j = 1,2,\ldots,n$, then we have:

- $\mathbb{E}(e_i) = 0$;
- $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$;
- $\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$.

Residuals

Let h_{ij} be the (i,j)th element of the matrix $\boldsymbol{H}$, $i,j = 1,2,\ldots,n$, then we have:

- $\mathbb{E}(e_i) = 0$;
- $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$;
- $\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$.

Estimation of the error variance

The estimator of the error variance, σ^2, is given by

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}} \right)^2}{n-p} = \frac{1}{n-p} \boldsymbol{e}^\top \boldsymbol{e} \tag{8}$$

$$= \frac{1}{n-p} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right).$$

Estimation of the error variance

The estimator of the error variance, σ^2, is given by

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}} \right)^2}{n-p} = \frac{1}{n-p} \boldsymbol{e}^\top \boldsymbol{e} \tag{8}$$
$$= \frac{1}{n-p} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right).$$

The estimator $\hat{\sigma}^2$, given in (8), is also called the **residual mean square** (MSRes).

Estimation of the error variance

The estimator of the error variance, σ^2, is given by

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}} \right)^2}{n-p} = \frac{1}{n-p} \boldsymbol{e}^\top \boldsymbol{e} \tag{8}$$
$$= \frac{1}{n-p} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right).$$

The estimator $\hat{\sigma}^2$, given in (8), is also called the **residual mean square** (MSRes). The quantity $\sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}} \right)^2$ is called the **residual sum of squares** (SSRes).

Estimation of the error variance

The estimator of the error variance, σ^2, is given by

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}} \right)^2}{n-p} = \frac{1}{n-p} \boldsymbol{e}^\top \boldsymbol{e} \tag{8}$$
$$= \frac{1}{n-p} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right).$$

The estimator $\hat{\sigma}^2$, given in (8), is also called the **residual mean square** (MSRes). The quantity $\sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}} \right)^2$ is called the **residual sum of squares** (SSRes).

Estimation of the error variance

The expectation of $\hat{\sigma}^2$ is given by

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2) &= \frac{1}{n-p}\mathbb{E}(\boldsymbol{e}^\top \boldsymbol{e}) = \frac{1}{n-p}\mathbb{E}\left\{[(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}]^\top (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}\right\} \\
&= \frac{1}{n-p}\mathbb{E}\left\{[(\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{X}\beta + \varepsilon)]^\top (\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{X}\beta + \varepsilon)\right\} \\
&\overset{*}{=} \frac{1}{n-p}\mathbb{E}\left\{[(\boldsymbol{I}_n - \boldsymbol{H})\varepsilon]^\top (\boldsymbol{I}_n - \boldsymbol{H})\varepsilon\right\} \\
&= \frac{1}{n-p}\mathbb{E}\left[\varepsilon^\top (\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{I}_n - \boldsymbol{H})\varepsilon\right] = \frac{1}{n-p}\mathbb{E}\left[\varepsilon^\top (\boldsymbol{I}_n - \boldsymbol{H})\varepsilon\right] \\
&\overset{**}{=} \frac{1}{n-p}\mathrm{tr}\left[(\boldsymbol{I}_n - \boldsymbol{H})\sigma^2 \boldsymbol{I}_n\right] = \frac{\sigma^2}{n-p}\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{H}) \\
&\overset{***}{=} \frac{\sigma^2}{n-p}(n-p) = \sigma^2.
\end{aligned}$$

Estimation of the error variance

The expectation of $\hat{\sigma}^2$ is given by

$$
\begin{aligned}
\mathbb{E}(\hat{\sigma}^2) &= \frac{1}{n-p}\mathbb{E}(\boldsymbol{e}^\top \boldsymbol{e}) = \frac{1}{n-p}\mathbb{E}\left\{[(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}]^\top (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}\right\} \\
&= \frac{1}{n-p}\mathbb{E}\left\{[(\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{X}\beta + \varepsilon)]^\top (\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{X}\beta + \varepsilon)\right\} \\
&\overset{*}{=} \frac{1}{n-p}\mathbb{E}\left\{[(\boldsymbol{I}_n - \boldsymbol{H})\varepsilon]^\top (\boldsymbol{I}_n - \boldsymbol{H})\varepsilon\right\} \\
&= \frac{1}{n-p}\mathbb{E}\left[\varepsilon^\top (\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{I}_n - \boldsymbol{H})\varepsilon\right] = \frac{1}{n-p}\mathbb{E}\left[\varepsilon^\top (\boldsymbol{I}_n - \boldsymbol{H})\varepsilon\right] \\
&\overset{**}{=} \frac{1}{n-p}\text{tr}\left[(\boldsymbol{I}_n - \boldsymbol{H})\sigma^2 \boldsymbol{I}_n\right] = \frac{\sigma^2}{n-p}\text{tr}(\boldsymbol{I}_n - \boldsymbol{H}) \\
&\overset{***}{=} \frac{\sigma^2}{n-p}(n-p) = \sigma^2.
\end{aligned}
$$

Estimation of the error variance

$*$: See $\mathbb{E}(\boldsymbol{e})$.

$**$: Let $\boldsymbol{v}$ be a random vector of length v, with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$, and let $\boldsymbol{W}$ be a matrix of dimension $v \times v$, then

Estimation of the error variance

$*$: See $\mathbb{E}(\boldsymbol{e})$.

$**$: Let $\boldsymbol{v}$ be a random vector of length v, with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$, and let $\boldsymbol{W}$ be a matrix of dimension $v \times v$, then

$$\mathbb{E}\left[\boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{v}\right] = \text{tr}\left[\boldsymbol{W}\boldsymbol{C}\right] + \boldsymbol{m}^\top \boldsymbol{W}\boldsymbol{m}.$$

Estimation of the error variance

$*$: See $\mathbb{E}(\boldsymbol{e})$.

$**$: Let $\boldsymbol{v}$ be a random vector of length v, with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$, and let $\boldsymbol{W}$ be a matrix of dimension $v \times v$, then

$$\mathbb{E}\left[\boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{v}\right] = \text{tr}\left[\boldsymbol{W}\boldsymbol{C}\right] + \boldsymbol{m}^\top \boldsymbol{W}\boldsymbol{m}.$$

$***$: $\text{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = \text{tr}(\boldsymbol{I}_n) - \text{tr}(\boldsymbol{H})$.

Estimation of the error variance

$*$: See $\mathbb{E}(\boldsymbol{e})$.

$**$: Let $\boldsymbol{v}$ be a random vector of length v, with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$, and let $\boldsymbol{W}$ be a matrix of dimension $v \times v$, then

$$\mathbb{E}\left[\boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{v}\right] = \text{tr}\left[\boldsymbol{W}\boldsymbol{C}\right] + \boldsymbol{m}^\top \boldsymbol{W}\boldsymbol{m}.$$

$***$: $\text{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = \text{tr}(\boldsymbol{I}_n) - \text{tr}(\boldsymbol{H})$. Since $\text{tr}(\boldsymbol{I}_n) = n$ and

$$\text{tr}(\boldsymbol{H}) = \text{tr}\left[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right] = \text{tr}\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right] = \text{tr}(\boldsymbol{I}_p) = p.$$

Estimation of the error variance

$*$: See $\mathbb{E}(\boldsymbol{e})$.

$**$: Let $\boldsymbol{v}$ be a random vector of length v, with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$, and let $\boldsymbol{W}$ be a matrix of dimension $v \times v$, then

$$\mathbb{E}\left[\boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{v}\right] = \text{tr}\left[\boldsymbol{W}\boldsymbol{C}\right] + \boldsymbol{m}^\top \boldsymbol{W}\boldsymbol{m}.$$

$***$: $\text{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = \text{tr}(\boldsymbol{I}_n) - \text{tr}(\boldsymbol{H})$. Since $\text{tr}(\boldsymbol{I}_n) = n$ and

$$\text{tr}(\boldsymbol{H}) = \text{tr}\left[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right] = \text{tr}\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right] = \text{tr}(\boldsymbol{I}_p) = p.$$

Therefore, $\text{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = n - p$.

Estimation of the error variance

$*$: See $\mathbb{E}(\boldsymbol{e})$.

$**$: Let $\boldsymbol{v}$ be a random vector of length v, with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$, and let $\boldsymbol{W}$ be a matrix of dimension $v \times v$, then

$$\mathbb{E}\left[\boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{v}\right] = \text{tr}\left[\boldsymbol{W}\boldsymbol{C}\right] + \boldsymbol{m}^\top \boldsymbol{W}\boldsymbol{m}.$$

$***$: $\text{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = \text{tr}(\boldsymbol{I}_n) - \text{tr}(\boldsymbol{H})$. Since $\text{tr}(\boldsymbol{I}_n) = n$ and

$$\text{tr}(\boldsymbol{H}) = \text{tr}\left[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right] = \text{tr}\left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right] = \text{tr}(\boldsymbol{I}_p) = p.$$

Therefore, $\text{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = n - p$.

Estimation of the error variance

And thus, we have that the MSRes, given in (8), is an unbiased estimator for σ^2.

Outline

Maximum likelihood method

Definition

The maximum likelihood estimators (MLEs) obtain estimates for unknown parameters that **maximize** the likelihood function.

Maximum likelihood method

Definition

The maximum likelihood estimators (MLEs) obtain estimates for unknown parameters that **maximize** the likelihood function.

Maximum likelihood method

Maximum likelihood (along with some of its variants) is the most widely used estimation method, and a list of its applications would cover practically the entire field of statistics (Lehmann and Casella, 1998, p. 468).

Maximum likelihood method

Maximum likelihood (along with some of its variants) is the most widely used estimation method, and a list of its applications would cover practically the entire field of statistics (Lehmann and Casella, 1998, p. 468).

The maximum likelihood method is by far the most popular technique for deriving estimators (Casella and Berger, 2002, p. 315).

Maximum likelihood method

Maximum likelihood (along with some of its variants) is the most widely used estimation method, and a list of its applications would cover practically the entire field of statistics (Lehmann and Casella, 1998, p. 468).

The maximum likelihood method is by far the most popular technique for deriving estimators (Casella and Berger, 2002, p. 315).

Maximum likelihood method

Let $Y_1, Y_2 \ldots, Y_n$ be independent random variables, with $Y_\ell \sim \mathcal{N}(\mu_\ell, \sigma^2)$,

Maximum likelihood method

Let $Y_1, Y_2 \ldots, Y_n$ be independent random variables, with $Y_\ell \sim \mathcal{N}(\mu_\ell, \sigma^2)$,

where $\mu_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta}$, $\ell = 1, 2, \ldots, n$,

Maximum likelihood method

Let $Y_1, Y_2 \ldots, Y_n$ be independent random variables, with $Y_\ell \sim \mathcal{N}(\mu_\ell, \sigma^2)$, where $\mu_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta}$, $\ell = 1, 2, \ldots, n$, the likelihood function is given by

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \boldsymbol{\beta} \right)^2 \right\}. \tag{9}$$

Maximum likelihood method

Let $Y_1, Y_2 \ldots, Y_n$ be independent random variables, with $Y_\ell \sim \mathcal{N}(\mu_\ell, \sigma^2)$, where $\mu_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta}$, $\ell = 1, 2, \ldots, n$, the likelihood function is given by

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \boldsymbol{\beta} \right)^2 \right\}. \tag{9}$$

Maximum likelihood method

Maximizing (9) is equivalent to maximizing its logarithm, given by

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\ell=1}^{n}\left(Y_\ell - \boldsymbol{x}_\ell^\top \boldsymbol{\beta}\right)^2. \tag{10}$$

Maximum likelihood method

Maximizing (9) is equivalent to maximizing its logarithm, given by

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\ell=1}^{n}\left(Y_\ell - \boldsymbol{x}_\ell^\top \boldsymbol{\beta}\right)^2. \tag{10}$$

Maximum likelihood method

Differentiating (10), with respect to the parameter β_m, $m = 1, 2, \ldots, p$, we have:

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_m} = \frac{1}{\sigma^2} \sum_{\ell=1}^{n} \left[Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p}) \right] x_{\ell m}. \qquad (11)$$

Maximum likelihood method

Differentiating (10), with respect to the parameter β_m, $m = 1, 2, \ldots, p$, we have:

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_m} = \frac{1}{\sigma^2} \sum_{\ell=1}^{n} \left[Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p}) \right] x_{\ell m}. \qquad (11)$$

Setting (11) equal to zero,

Maximum likelihood method

Differentiating (10), with respect to the parameter β_m, $m = 1, 2, \ldots, p$, we have:

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_m} = \frac{1}{\sigma^2} \sum_{\ell=1}^{n} \left[Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p}) \right] x_{\ell m}. \qquad (11)$$

Setting (11) equal to zero, we have

$$\hat{\beta}_1 \sum_{\ell=1}^{n} x_{\ell 1} x_{\ell m} + \cdots + \hat{\beta}_m \sum_{\ell=1}^{n} x_{\ell m}^2 + \cdots + \hat{\beta}_p \sum_{\ell=1}^{n} x_{\ell p} x_{\ell m} = \sum_{\ell=1}^{n} x_{\ell m} Y_\ell. \quad (12)$$

Maximum likelihood method

Differentiating (10), with respect to the parameter β_m, $m = 1, 2, \ldots, p$, we have:

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_m} = \frac{1}{\sigma^2} \sum_{\ell=1}^{n} \left[Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p}) \right] x_{\ell m}. \qquad (11)$$

Setting (11) equal to zero, we have

$$\hat{\beta}_1 \sum_{\ell=1}^{n} x_{\ell 1} x_{\ell m} + \cdots + \hat{\beta}_m \sum_{\ell=1}^{n} x_{\ell m}^2 + \cdots + \hat{\beta}_p \sum_{\ell=1}^{n} x_{\ell p} x_{\ell m} = \sum_{\ell=1}^{n} x_{\ell m} Y_\ell. \quad (12)$$

Maximum likelihood method

The **maximum likelihood estimator** of β_m, obtained from the solution of the p equations in (12), is given by:

$$\widehat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}. \tag{13}$$

Maximum likelihood method

The **maximum likelihood estimator** of β_m, obtained from the solution of the p equations in (12), is given by:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}. \tag{13}$$

The MLE in (13) coincides with the LSE, presented in (6).

Maximum likelihood method

The **maximum likelihood estimator** of β_m, obtained from the solution of the p equations in (12), is given by:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}. \tag{13}$$

The MLE in (13) coincides with the LSE, presented in (6).

Maximum likelihood method

Now, differentiating (10), with respect to the parameter σ^2, we have:

$$\frac{\partial l(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \beta \right)^2 . \tag{14}$$

Maximum likelihood method

Now, differentiating (10), with respect to the parameter σ^2, we have:

$$\frac{\partial l(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \beta \right)^2 . \tag{14}$$

Setting (14) equal to zero,

Maximum likelihood method

Now, differentiating (10), with respect to the parameter σ^2, we have:

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \boldsymbol{\beta} \right)^2. \tag{14}$$

Setting (14) equal to zero, we have

$$\frac{1}{2(\hat{\sigma}^2)^2} \sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}} \right)^2 = \frac{n}{2\hat{\sigma}^2}. \tag{15}$$

Maximum likelihood method

Now, differentiating (10), with respect to the parameter σ^2, we have:

$$\frac{\partial l(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \beta \right)^2. \tag{14}$$

Setting (14) equal to zero, we have

$$\frac{1}{2(\hat{\sigma}^2)^2} \sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\beta} \right)^2 = \frac{n}{2\hat{\sigma}^2}. \tag{15}$$

Maximum likelihood method

The **maximum likelihood estimator** of σ^2, obtained from the solution of equation (15), is given by:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}} \right)^2}{n} = \left(\frac{n-p}{n} \right) \text{MSRes.} \tag{16}$$

Maximum likelihood method

The **maximum likelihood estimator** of σ^2, obtained from the solution of equation (15), is given by:

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}} \right)^2}{n} = \left(\frac{n-p}{n} \right) \mathrm{MSRes}. \qquad (16)$$

Maximum likelihood method

The estimator for $\boldsymbol{\beta}$ obtained by the maximum likelihood method coincides with that obtained by the least squares method, therefore, the MLE of $\boldsymbol{\beta}$ is **unbiased** and has the smallest variance among all linear unbiased estimators.

Maximum likelihood method

The estimator for $\boldsymbol{\beta}$ obtained by the maximum likelihood method coincides with that obtained by the least squares method, therefore, the MLE of $\boldsymbol{\beta}$ is **unbiased** and has the smallest variance among all linear unbiased estimators.

Maximum likelihood method

The same does not hold for the MLE of σ^2. In fact, $\hat{\sigma}^2_{\mathrm{MLE}}$ is a **biased estimator**,

Maximum likelihood method

The same does not hold for the MLE of σ^2. In fact, $\hat{\sigma}^2_{\mathrm{MLE}}$ is a **biased estimator**, and for small sample sizes, this bias is considerable.

Maximum likelihood method

The same does not hold for the MLE of σ^2. In fact, $\hat{\sigma}^2_{\mathrm{MLE}}$ is a **biased estimator**, and for small sample sizes, this bias is considerable.

Additionally, the MLEs of β and σ^2 are consistent estimators.

Maximum likelihood method

The same does not hold for the MLE of σ^2. In fact, $\hat{\sigma}^2_{\mathrm{MLE}}$ is a **biased estimator**, and for small sample sizes, this bias is considerable.

Additionally, the MLEs of $\boldsymbol{\beta}$ and σ^2 are consistent estimators.

The normality assumption and the resulting maximum likelihood estimation allow us to construct inferential procedures, such as confidence intervals.

Maximum likelihood method

The same does not hold for the MLE of σ^2. In fact, $\hat{\sigma}^2_{\mathrm{MLE}}$ is a **biased estimator**, and for small sample sizes, this bias is considerable.

Additionally, the MLEs of $\boldsymbol{\beta}$ and σ^2 are consistent estimators.

The normality assumption and the resulting maximum likelihood estimation allow us to construct inferential procedures, such as confidence intervals.

Maximum likelihood method

Differentiating (10) a second time, with respect to the parameter $\beta_{m'}$, $m' = 1, 2, \ldots, p$, we have:

$$\frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_m \partial \beta_{m'}} = -\frac{1}{\sigma^2} \sum_{\ell=1}^{n} x_{\ell m} x_{\ell m'}. \tag{17}$$

Maximum likelihood method

Differentiating (10) a second time, with respect to the parameter $\beta_{m'}$, $m' = 1, 2, \ldots, p$, we have:

$$\frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_m \partial \beta_{m'}} = -\frac{1}{\sigma^2} \sum_{\ell=1}^{n} x_{\ell m} x_{\ell m'}. \tag{17}$$

If we take the expectation and multiply by -1 in (17), we obtain the (m, m')th element of the Fisher information matrix for $\boldsymbol{\beta}$.

Maximum likelihood method

Differentiating (10) a second time, with respect to the parameter $\beta_{m'}$, $m' = 1, 2, \ldots, p$, we have:

$$\frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_m \partial \beta_{m'}} = -\frac{1}{\sigma^2} \sum_{\ell=1}^{n} x_{\ell m} x_{\ell m'}. \tag{17}$$

If we take the expectation and multiply by -1 in (17), we obtain the (m, m')th element of the Fisher information matrix for $\boldsymbol{\beta}$.

Maximum likelihood method

Recall that equation (11) is the **score function**. In matrix notation, the score vector and the Fisher information matrix are given, respectively, by

Maximum likelihood method

Recall that equation (11) is the **score function**. In matrix notation, the score vector and the Fisher information matrix are given, respectively, by

$$\boldsymbol{U}_\beta = \frac{1}{\sigma^2}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\beta) \text{ e } \boldsymbol{K}_\beta = \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X}. \tag{18}$$

Maximum likelihood method

Recall that equation (11) is the **score function**. In matrix notation, the score vector and the Fisher information matrix are given, respectively, by

$$\boldsymbol{U}_\beta = \frac{1}{\sigma^2}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\beta) \text{ e } \boldsymbol{K}_\beta = \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X}. \tag{18}$$

Distribution of the MLE

Under the normal linear model, the MLE of β has the exact distribution

$$\widehat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right).$$

Consequently, for a linear combination $\boldsymbol{a}^\top \beta$,

$$\boldsymbol{a}^\top \widehat{\beta} \sim \mathcal{N}\left(\boldsymbol{a}^\top \beta, \sigma^2 \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}\right).$$

Distribution of the MLE

Under the normal linear model, the MLE of β has the exact distribution

$$\widehat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right).$$

Consequently, for a linear combination $\boldsymbol{a}^\top \beta$,

$$\boldsymbol{a}^\top \widehat{\beta} \sim \mathcal{N}\left(\boldsymbol{a}^\top \beta, \sigma^2 \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}\right).$$

This result provides the basis for constructing confidence intervals and hypothesis tests for linear combinations of the regression parameters.

Distribution of the MLE

Under the normal linear model, the MLE of β has the exact distribution

$$\widehat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right).$$

Consequently, for a linear combination $\boldsymbol{a}^\top \beta$,

$$\boldsymbol{a}^\top \widehat{\beta} \sim \mathcal{N}\left(\boldsymbol{a}^\top \beta, \sigma^2 \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}\right).$$

This result provides the basis for constructing confidence intervals and hypothesis tests for linear combinations of the regression parameters.

Distribution of the MLE

For the MLE of σ^2, we have

$$\frac{n\hat{\sigma}^2_{\text{MLE}}}{\sigma^2} \sim \chi^2_{n-p}.$$

Distribution of the MLE

For the MLE of σ^2, we have

$$\frac{n\hat{\sigma}^2_{\mathrm{MLE}}}{\sigma^2} \sim \chi^2_{n-p}.$$

Bias of the MLE of σ^2

From the previous result,

$$\mathbb{E}(\hat{\sigma}^2_{\mathrm{MLE}}) = \left(\frac{n-p}{n}\right)\sigma^2.$$

Therefore, $\hat{\sigma}^2_{\mathrm{MLE}}$ is biased,

Bias of the MLE of σ^2

From the previous result,

$$\mathbb{E}(\hat{\sigma}^2_{\mathrm{MLE}}) = \left(\frac{n-p}{n}\right)\sigma^2.$$

Therefore, $\hat{\sigma}^2_{\mathrm{MLE}}$ is biased, with

$$\mathrm{Bias}(\hat{\sigma}^2_{\mathrm{MLE}}) = -\frac{p\sigma^2}{n}.$$

Bias of the MLE of σ^2

From the previous result,

$$\mathbb{E}(\hat{\sigma}^2_{\mathrm{MLE}}) = \left(\frac{n-p}{n}\right)\sigma^2.$$

Therefore, $\hat{\sigma}^2_{\mathrm{MLE}}$ is biased, with

$$\mathrm{Bias}(\hat{\sigma}^2_{\mathrm{MLE}}) = -\frac{p\sigma^2}{n}.$$

The bias vanishes as $n \to \infty$.

Bias of the MLE of σ^2

From the previous result,

$$\mathbb{E}(\hat{\sigma}^2_{\mathrm{MLE}}) = \left(\frac{n-p}{n}\right)\sigma^2.$$

Therefore, $\hat{\sigma}^2_{\mathrm{MLE}}$ is biased, with

$$\mathrm{Bias}(\hat{\sigma}^2_{\mathrm{MLE}}) = -\frac{p\sigma^2}{n}.$$

The bias vanishes as $n \to \infty$.

Considerations

Although it is common to use $\hat{Y}_\ell = \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}}$,

Considerations

Although it is common to use $\hat{Y}_\ell = \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}}$, what we are actually estimating is μ_ℓ.

Considerations

Although it is common to use $\hat{Y}_\ell = \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}}$, what we are actually estimating is μ_ℓ. Recall that $\mu_\ell = \mathbb{E}(Y_\ell) = \boldsymbol{x}_\ell^\top \boldsymbol{\beta}$.

Considerations

Although it is common to use $\hat{Y}_\ell = \boldsymbol{x}_\ell^\top \widehat{\boldsymbol{\beta}}$, what we are actually estimating is μ_ℓ. Recall that $\mu_\ell = \mathbb{E}(Y_\ell) = \boldsymbol{x}_\ell^\top \boldsymbol{\beta}$.

Outline

Gradient descent

In the linear regression model, the least squares estimator has a closed-form solution. However, in many statistical and machine learning models, closed-form solutions are not available.

Gradient descent

In the linear regression model, the least squares estimator has a closed-form solution. However, in many statistical and machine learning models, closed-form solutions are not available.

In such cases, **iterative optimization methods** can be used to obtain parameter estimates.

Gradient descent

In the linear regression model, the least squares estimator has a closed-form solution. However, in many statistical and machine learning models, closed-form solutions are not available.

In such cases, **iterative optimization methods** can be used to obtain parameter estimates. One of the most fundamental examples is the **gradient descent** method.

Gradient descent

In the linear regression model, the least squares estimator has a closed-form solution. However, in many statistical and machine learning models, closed-form solutions are not available.

In such cases, **iterative optimization methods** can be used to obtain parameter estimates. One of the most fundamental examples is the **gradient descent** method.

Gradient descent for least squares

Consider the residual sum of squares, defined in (3), as the cost function.

Its gradient, $\nabla Q(\beta)$, is given in (7).

Gradient descent for least squares

Consider the residual sum of squares, defined in (3), as the cost function.

Its gradient, $\nabla Q(\beta)$, is given in (7). Starting from an initial value $\beta^{(0)}$, gradient descent iteratively updates the parameters:

$$\beta^{(k+1)} = \beta^{(k)} - \text{LR} \nabla Q(\beta^{(k)}),$$

where LR is the *learning rate*.

Gradient descent for least squares

Consider the residual sum of squares, defined in (3), as the cost function. Its gradient, $\nabla Q(\beta)$, is given in (7). Starting from an initial value $\beta^{(0)}$, gradient descent iteratively updates the parameters:

$$\beta^{(k+1)} = \beta^{(k)} - \text{LR}\,\nabla Q(\beta^{(k)}),$$

where LR is the *learning rate*.

Gradient descent for least squares

Unlike the normal equations, gradient descent does not require computing $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

Each iteration has computational cost $\mathcal{O}(np)$, so the total cost is $\mathcal{O}(Knp)$ for K iterations.

Gradient descent for least squares

Unlike the normal equations, gradient descent does not require computing $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

Each iteration has computational cost $\mathcal{O}(np)$, so the total cost is $\mathcal{O}(Knp)$ for K iterations.

The number of iterations required for convergence depends, among other factors, on the choice of the learning rate.

Gradient descent for least squares

Unlike the normal equations, gradient descent does not require computing $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

Each iteration has computational cost $\mathcal{O}(np)$, so the total cost is $\mathcal{O}(Knp)$ for K iterations.

The number of iterations required for convergence depends, among other factors, on the choice of the learning rate.

Outline

Applications

Application 1. *Westwood Company* dataset (Neter et al., 1983, p. 36).

After fitting the model (estimating the parameters), we obtain the following fitted model,

$$\hat{Y}_\ell = 10 + 2x_{\ell 2},$$

Applications

Application 1. *Westwood Company* dataset (Neter et al., 1983, p. 36). After fitting the model (estimating the parameters), we obtain the following fitted model,

$$\hat{Y}_\ell = 10 + 2x_{\ell 2},$$

where Y_ℓ: man-hours, $x_{\ell 2}$: lot size, $\ell = 1, 2, \ldots, 10$, and MSRes $= 7.5$.

Applications

Application 1. *Westwood Company* dataset (Neter et al., 1983, p. 36). After fitting the model (estimating the parameters), we obtain the following fitted model,

$$\hat{Y}_\ell = 10 + 2x_{\ell 2},$$

where Y_ℓ: man-hours, $x_{\ell 2}$: lot size, $\ell = 1, 2, \ldots, 10$, and MSRes $= 7.5$.

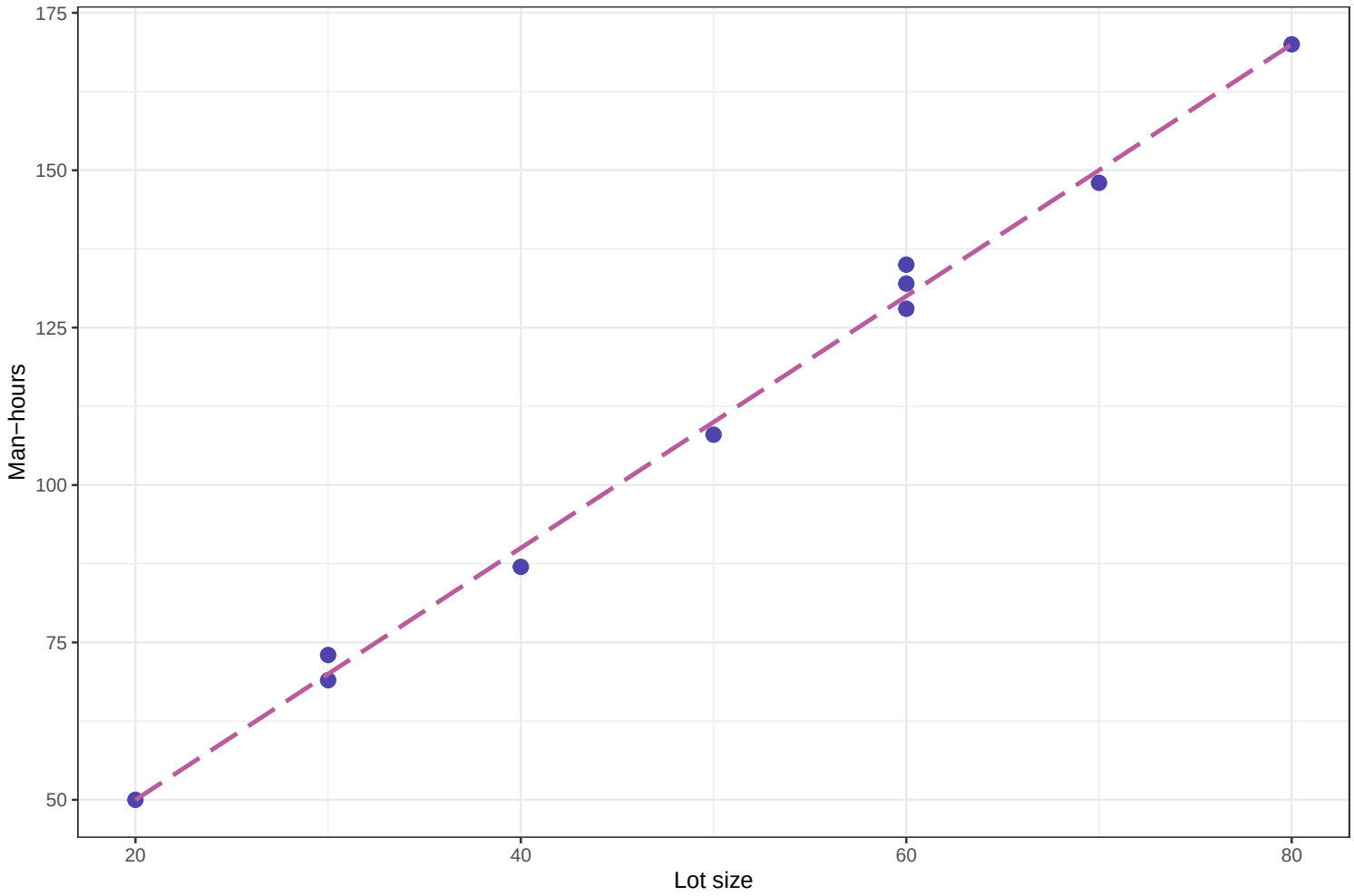

Figure 1: Scatterplot with the fitted line for the *Westwood Company* data.

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = 10$:

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = 10$: if the lot size were zero, the expected number of man-hours would be 10;

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = 10$: if the lot size were zero, the expected number of man-hours would be 10;
- $\hat{\beta}_2 = 2$:

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = 10$: if the lot size were zero, the expected number of man-hours would be 10;
- $\hat{\beta}_2 = 2$: for each one-unit increase in lot size, the expected number of man-hours increases by 2.

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = 10$: if the lot size were zero, the expected number of man-hours would be 10;
- $\hat{\beta}_2 = 2$: for each one-unit increase in lot size, the expected number of man-hours increases by 2.

Applications

Application 2. Dataset from Walker (1962). The original temperature measurements are reported in °F; here, we use the equivalent measurements in °C (McDonald, 2009). After fitting the model, we have the following estimated model,

$$\hat{Y}_\ell = -7.211 + 3.603 x_{\ell 2} - 10.065 x_{\ell 3},$$

Applications

Application 2. Dataset from Walker (1962). The original temperature measurements are reported in °F; here, we use the equivalent measurements in °C (McDonald, 2009). After fitting the model, we have the following estimated model,

$$\hat{Y}_\ell = -7.211 + 3.603x_{\ell 2} - 10.065x_{\ell 3},$$

where Y_ℓ: number of pulses per second, $x_{\ell 2}$: temperature (°C), $x_{\ell 3}$: cricket species, $x_{\ell 3} \in \{Oecanthus\ exclamationis\ (ex), Oecanthus\ niveus\ (niv)\}$, $\ell = 1, 2, \ldots, 31$ and MSRes $= 12.764$.

Applications

Application 2. Dataset from Walker (1962). The original temperature measurements are reported in °F; here, we use the equivalent measurements in °C (McDonald, 2009). After fitting the model, we have the following estimated model,

$$\hat{Y}_\ell = -7.211 + 3.603 x_{\ell 2} - 10.065 x_{\ell 3},$$

where Y_ℓ: number of pulses per second, $x_{\ell 2}$: temperature (°C), $x_{\ell 3}$: cricket species, $x_{\ell 3} \in \{\textit{Oecanthus exclamationis (ex)}, \textit{Oecanthus niveus (niv)}\}$, $\ell = 1, 2, \ldots, 31$ and MSRes $= 12.764$.

Applications

An alternative way to present the fitted model is as follows:

$$\begin{cases} \textit{Oecanthus niveus}: & \hat{Y}_\ell = -17.276 + 3.603 x_{\ell 2} \\ \textit{Oecanthus exclamationis}: & \hat{Y}_\ell = -7.211 + 3.603 x_{\ell 2} \end{cases}.$$

Applications

An alternative way to present the fitted model is as follows:

$$\begin{cases} \textit{Oecanthus niveus}: & \hat{Y}_\ell = -17.276 + 3.603 x_{\ell 2} \\ \textit{Oecanthus exclamationis}: & \hat{Y}_\ell = -7.211 + 3.603 x_{\ell 2} \end{cases}.$$

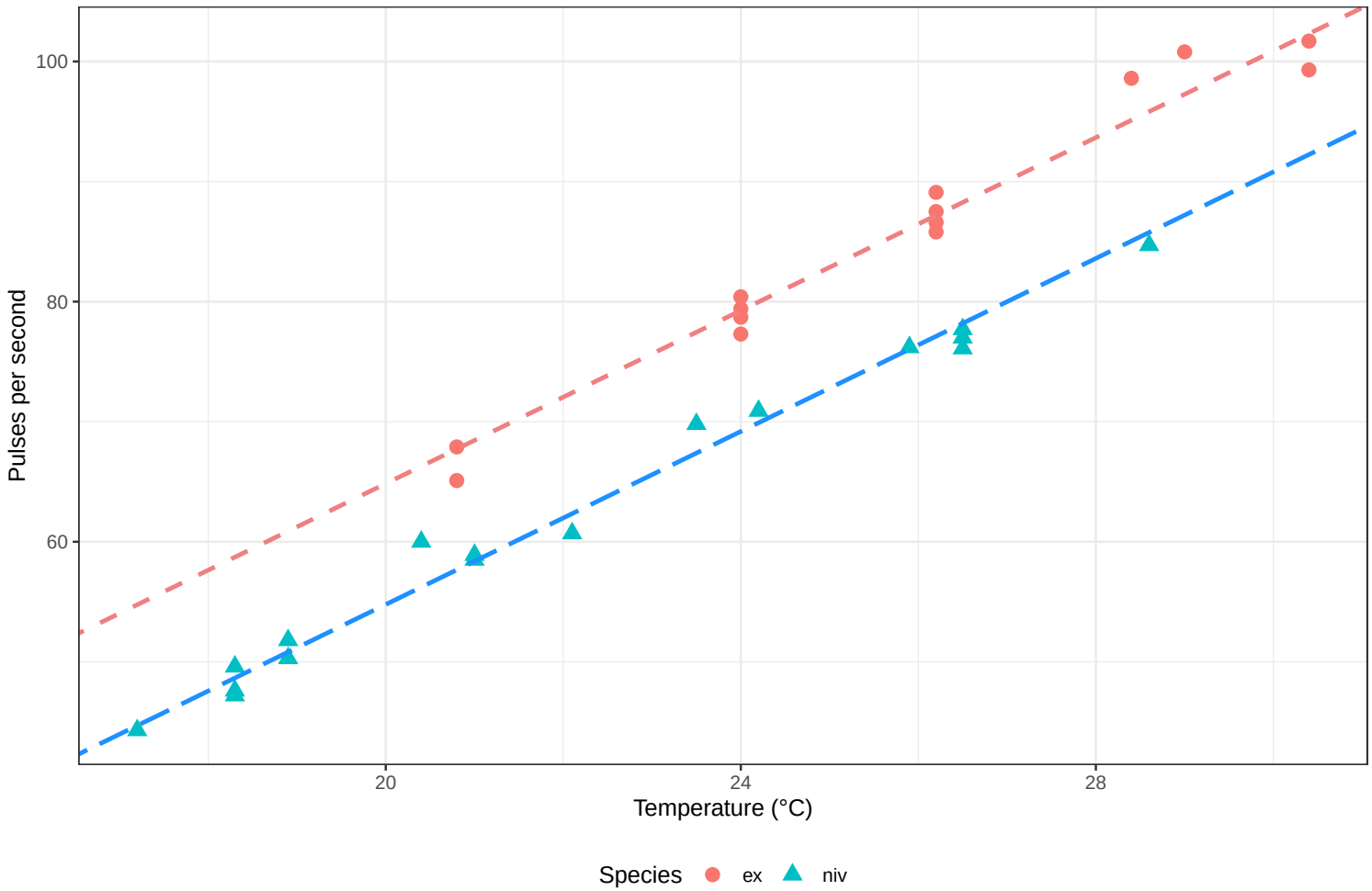

Figure 2: Scatterplot with the fitted lines for the data from Walker (1962).

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = -7.211$:

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = -7.211$: if the temperature were zero and the cricket species were *Oecanthus exclamationis*, the expected number of pulses per second would be -7.211 (does this make sense?);

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = -7.211$: if the temperature were zero and the cricket species were *Oecanthus exclamationis*, the expected number of pulses per second would be -7.211 (does this make sense?);
- $\hat{\beta}_2 = 3.603$:

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = -7.211$: if the temperature were zero and the cricket species were *Oecanthus exclamationis*, the expected number of pulses per second would be -7.211 (does this make sense?);
- $\hat{\beta}_2 = 3.603$: for each one-unit increase in temperature, the expected number of pulses per second increases by 3.603;

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = -7.211$: if the temperature were zero and the cricket species were *Oecanthus exclamationis*, the expected number of pulses per second would be -7.211 (does this make sense?);
- $\hat{\beta}_2 = 3.603$: for each one-unit increase in temperature, the expected number of pulses per second increases by 3.603;
- $\hat{\beta}_3 = -10.065$:

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = -7.211$: if the temperature were zero and the cricket species were *Oecanthus exclamationis*, the expected number of pulses per second would be -7.211 (does this make sense?);
- $\hat{\beta}_2 = 3.603$: for each one-unit increase in temperature, the expected number of pulses per second increases by 3.603;
- $\hat{\beta}_3 = -10.065$: *Oecanthus niveus* has, on average, 10.065 fewer pulses per second than *Oecanthus exclamationis*, at the same temperature.

Applications

The parameters can be interpreted as follows:

- $\hat{\beta}_1 = -7.211$: if the temperature were zero and the cricket species were *Oecanthus exclamationis*, the expected number of pulses per second would be -7.211 (does this make sense?);
- $\hat{\beta}_2 = 3.603$: for each one-unit increase in temperature, the expected number of pulses per second increases by 3.603;
- $\hat{\beta}_3 = -10.065$: *Oecanthus niveus* has, on average, 10.065 fewer pulses per second than *Oecanthus exclamationis*, at the same temperature.

Outline

Simulation study

Consider the following NLM,

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

Simulation study

Consider 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

Consider the following NLM,

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

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

Simulation study

Consider the following NLM,

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

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

Simulation study

We will conduct a Monte Carlo study with 5,000 replications, where we estimate the absolute bias of the estimators of the parameters β_1, β_2, and σ^2,

Simulation study

We will conduct a Monte Carlo study with 5,000 replications, where we estimate the absolute bias of the estimators of the parameters β_1, β_2, and σ^2, for the latter, using the MSRes and the MLE,

Simulation study

We will conduct a Monte Carlo study with 5,000 replications, where we estimate the absolute bias of the estimators of the parameters β_1, β_2, and σ^2, for the latter, using the MSRes and the MLE, for $n \in \{10, 20, 40, 80, 160\}$.

Simulation study

We will conduct a Monte Carlo study with 5,000 replications, where we estimate the absolute bias of the estimators of the parameters β_1, β_2, and σ^2, for the latter, using the MSRes and the MLE, for $n \in \{10, 20, 40, 80, 160\}$.

Without loss of generality, the estimated absolute bias for the estimator of θ is given by

$$\text{Absolute bias} = \left|\hat{\theta}_{\text{MC}} - \theta\right|,$$

Simulation study

We will conduct a Monte Carlo study with 5,000 replications, where we estimate the absolute bias of the estimators of the parameters β_1, β_2, and σ^2, for the latter, using the MSRes and the MLE, for $n \in \{10, 20, 40, 80, 160\}$. Without loss of generality, the estimated absolute bias for the estimator of θ is given by

$$\text{Absolute bias} = \left|\hat{\theta}_{\text{MC}} - \theta\right|,$$

where $\hat{\theta}_{\text{MC}} = 5{,}000^{-1} \sum_{i=1}^{5{,}000} \hat{\theta}_i$.

Simulation study

We will conduct a Monte Carlo study with 5,000 replications, where we estimate the absolute bias of the estimators of the parameters β_1, β_2, and σ^2, for the latter, using the MSRes and the MLE, for $n \in \{10, 20, 40, 80, 160\}$. Without loss of generality, the estimated absolute bias for the estimator of θ is given by

$$\text{Absolute bias} = \left|\hat{\theta}_{\text{MC}} - \theta\right|,$$

where $\hat{\theta}_{\text{MC}} = 5{,}000^{-1} \sum_{i=1}^{5{,}000} \hat{\theta}_i$.

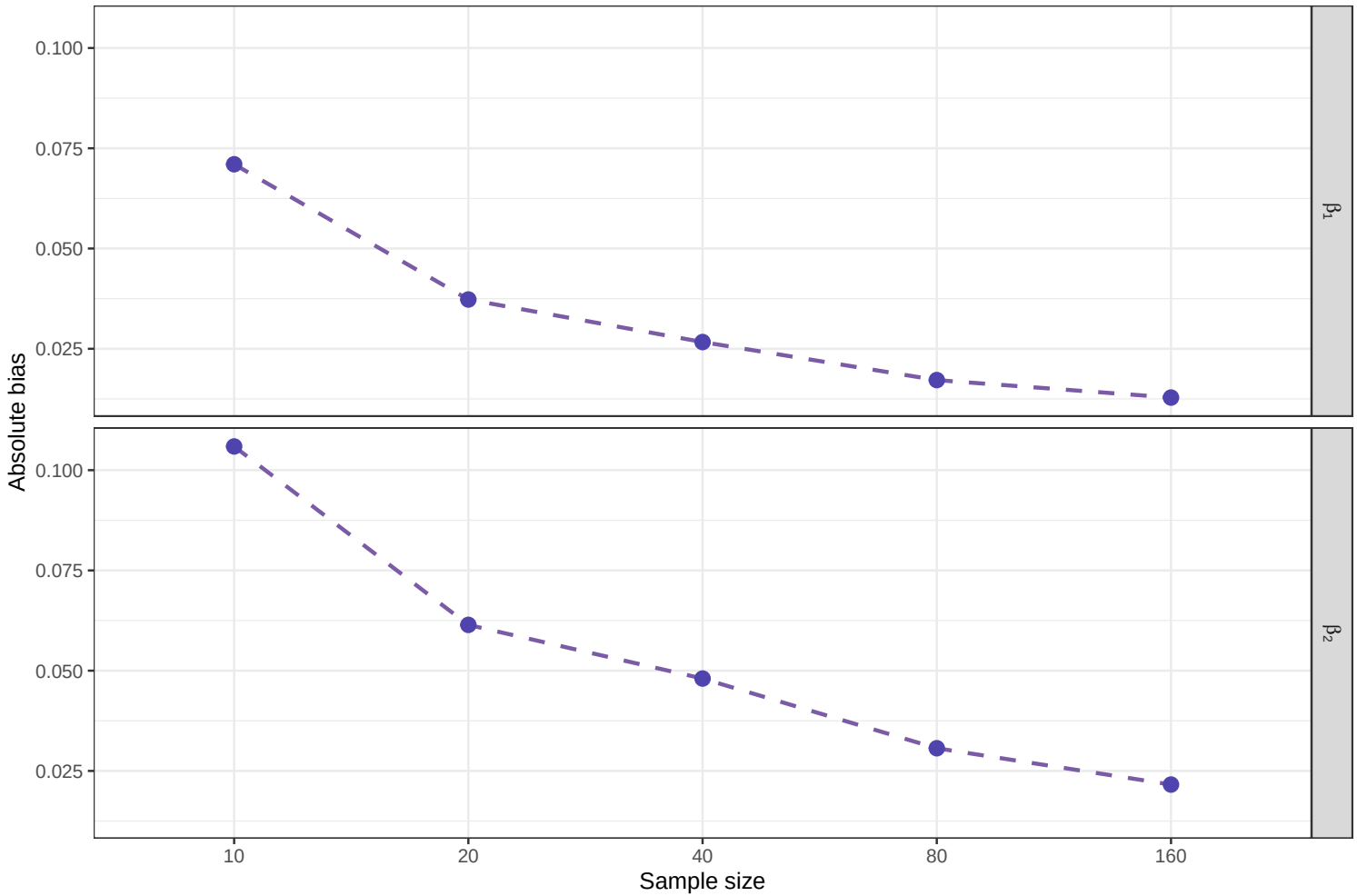

Figure 3: Absolute bias for β.

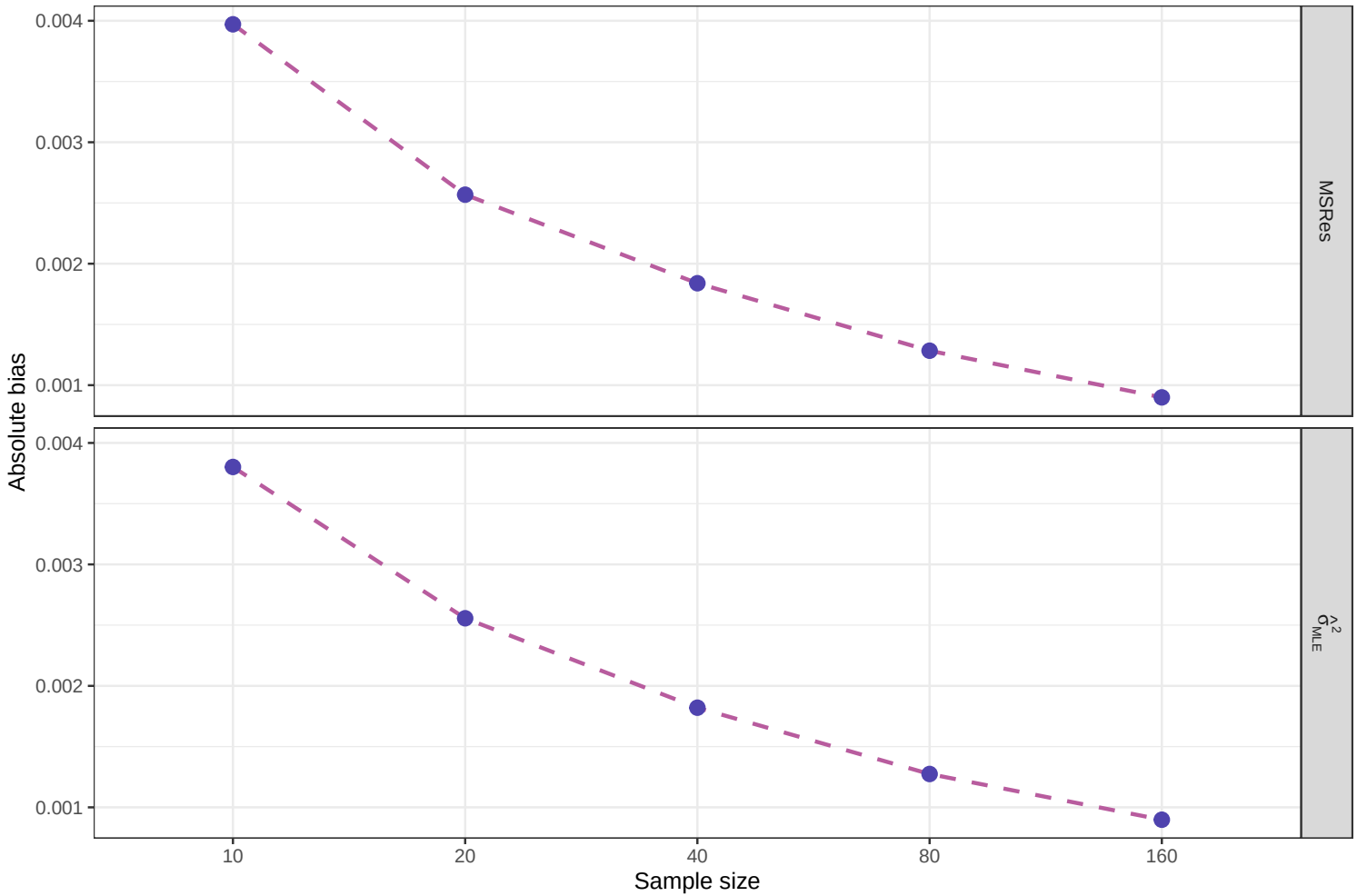

Figure 4: Absolute bias for σ^2, using the MSRes and the MLE.

Simulation study

Comments

From Figures 3 and 4, we can see that, as the sample size increases, the absolute bias converges to zero.

Simulation study

Comments

From Figures 3 and 4, we can see that, as the sample size increases, the absolute bias converges to zero.

Outline

References

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury/Thomson Learning, Pacific Grove, CA, USA, 2nd edition.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York, NY, USA, 2nd edition.

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

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

References

Plackett, R. L. (1949). A historical note on the method of least squares. *Biometrika*, 36(3/4):458–460.

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