Definições e suposições

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 18 de março de 2024

Roteiro

Roteiro

Introdução

Os primeiros conceitos de regressão foram propostos por Galton (1889), aplicados na Antropometria, em que o objetivo era equacionar as relações de dependência entre a altura dos pais (**variável explicativa** ou covariável)

Introdução

Os primeiros conceitos de regressão foram propostos por Galton (1889), aplicados na Antropometria, em que o objetivo era equacionar as relações de dependência entre a altura dos pais (**variável explicativa** ou covariável) e a altura dos filhos (variável de interesse ou **variável resposta** ou desfecho).

Introdução

Os primeiros conceitos de regressão foram propostos por Galton (1889), aplicados na Antropometria, em que o objetivo era equacionar as relações de dependência entre a altura dos pais (**variável explicativa** ou covariável) e a altura dos filhos (variável de interesse ou **variável resposta** ou desfecho).

Introdução

Porém, esses conceitos podem ser aplicados em qualquer contexto, como:

- na necessidade de uma empresa em analisar os fatores que podem interferir nas vendas;

Introdução

Porém, esses conceitos podem ser aplicados em qualquer contexto, como:

- na necessidade de uma empresa em analisar os fatores que podem interferir nas vendas;
- na predição de mortalidade de uma pessoa ao contrair uma doença, baseada nas suas características, como idade e doenças pré existentes.

Porém, esses conceitos podem ser aplicados em qualquer contexto, como:

- na necessidade de uma empresa em analisar os fatores que podem interferir nas vendas;
- na predição de mortalidade de uma pessoa ao contrair uma doença, baseada nas suas características, como idade e doenças pré existentes.

Introdução

A **análise de regressão** é uma técnica estatística para investigar e modelar a relação entre as variáveis.

Na verdade, a análise de regressão pode ser a técnica estatística mais utilizada (Montgomery et al., 2012).

Introdução

A **análise de regressão** é uma técnica estatística para investigar e modelar a relação entre as variáveis.

Na verdade, a análise de regressão pode ser a técnica estatística mais utilizada (Montgomery et al., 2012).

Roteiro

Modelo de regressão linear

Suponham que $Y_1, Y_2, \ldots, Y_n$ sejam variáveis aleatórias independentes, com $Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2)$, tal que

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

Modelo de regressão linear

Suponham que $Y_1, Y_2, \ldots, Y_n$ sejam variáveis aleatórias independentes, com $Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2)$, tal que

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

em que $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ são variáveis (quantidades) conhecidas, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ são parâmetros desconhecidos a serem estimados e σ^2 é a variância de Y_ℓ, $\ell = 1, 2, \ldots, n$, também desconhecida, a ser estimada.

Modelo de regressão linear

Suponham que $Y_1, Y_2, \ldots, Y_n$ sejam variáveis aleatórias independentes, com $Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2)$, tal que

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

em que $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ são variáveis (quantidades) conhecidas, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ são parâmetros desconhecidos a serem estimados e σ^2 é a variância de Y_ℓ, $\ell = 1, 2, \ldots, n$, também desconhecida, a ser estimada.

Modelo de regressão linear

Nós podemos admitir que

$$
\begin{aligned}
Y_\ell &= \beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p} + \varepsilon_\ell, \qquad (1)\\
&= \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell,
\end{aligned}
$$

Modelo de regressão linear

Nós podemos admitir que

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

em que $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2.

Modelo de regressão linear

Nós podemos admitir que

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

em que $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2.

Modelo de regressão linear

A Equação (1) é que nós definimos como **modelo de regressão linear** (MRL).

Observação: Muitos autores não se preocupam (diretamente) com a distribuição de Y_ℓ, e sim, com a de ε_ℓ, $\ell = 1, 2, \ldots, n$.

Modelo de regressão linear

A Equação (1) é que nós definimos como **modelo de regressão linear** (MRL).

Observação: Muitos autores não se preocupam (diretamente) com a distribuição de Y_ℓ, e sim, com a de ε_ℓ, $\ell = 1, 2, \ldots, n$.

Interpretação dos parâmetros do modelo

No MRL, os vetores $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ e $\boldsymbol{x}_\ell$, $\ell = 1, 2, \ldots, n$, são chamados de vetor resposta e preditor, respectivamente, e os parâmetros pertencentes ao vetor β são chamados de coeficientes de regressão.

Interpretação dos parâmetros do modelo

No MRL, os vetores $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ e $\boldsymbol{x}_\ell$, $\ell = 1, 2, \ldots, n$, são chamados de vetor resposta e preditor, respectivamente, e os parâmetros pertencentes ao vetor $\boldsymbol{\beta}$ são chamados de coeficientes de regressão.

Interpretação dos parâmetros do modelo

Parâmetro β_m, $m = 1, 2, \ldots, p$

Ele é o acréscimo esperado na variável resposta, quando nós aumentamos uma unidade em x_m e mantemos as demais variáveis preditoras constantes.

Interpretação dos parâmetros do modelo

Parâmetro β_m, $m = 1, 2, \ldots, p$

Ele é o acréscimo esperado na variável resposta, quando nós aumentamos uma unidade em x_m e mantemos as demais variáveis preditoras constantes.

Interpretação dos parâmetros do modelo

Intercepto

Ele é o valor esperado da variável resposta quando todas as $p-1$ variáveis preditoras valem 0.

Interpretação dos parâmetros do modelo

Intercepto

Ele é o valor esperado da variável resposta quando todas as $p - 1$ variáveis preditoras valem 0.

Observação: Para que (1) seja um MRL com intercepto, basta nós assumirmos que $x_{\ell 1} = 1, \forall \ell = 1, 2, \ldots, n$.

Interpretação dos parâmetros do modelo

Intercepto

Ele é o valor esperado da variável resposta quando todas as $p - 1$ variáveis preditoras valem 0.

Observação: Para que (1) seja um MRL com intercepto, basta nós assumirmos que $x_{\ell 1} = 1$, $\forall \ell = 1, 2, \ldots, n$.

Modelo de regressão linear simples

Quando um MRL pode ser escrito da seguinte forma:

Modelo de regressão linear simples

Quando um MRL pode ser escrito da seguinte forma:

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

Modelo de regressão linear simples

Quando um MRL pode ser escrito da seguinte forma:

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

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

Modelo de regressão linear simples

Quando um MRL pode ser escrito da seguinte forma:

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

em que $\boldsymbol{x}_\ell = (1, x_{\ell 2})^\top$, $\ell = 1, 2, \ldots, n$ e $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$. Nós denominamos (2) como **modelo de regressão linear simples** (MLS).

Modelo de regressão linear simples

Quando um MRL pode ser escrito da seguinte forma:

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

em que $\boldsymbol{x}_\ell = (1, x_{\ell 2})^\top$, $\ell = 1, 2, \ldots, n$ e $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$. Nós denominamos (2) como **modelo de regressão linear simples** (MLS).

Forma matricial

Um MRL pode ser escrito de forma matricial,

Forma matricial

Um MRL pode ser escrito de forma matricial, da seguinte forma:

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

Forma matricial

Um MRL pode ser escrito de forma matricial, da seguinte forma:

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

em que $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, sendo $\mathbf{0}$ o vetor nulo de dimensão n, $\boldsymbol{I}_n$ a matriz identidade de ordem n e $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ ou

Forma matricial

Um MRL pode ser escrito de forma matricial, da seguinte forma:

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

em que $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, sendo $\mathbf{0}$ o vetor nulo de dimensão n, $\boldsymbol{I}_n$ a matriz identidade de ordem n e $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ ou

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

Forma matricial

Um MRL pode ser escrito de forma matricial, da seguinte forma:

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

em que $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{D}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, sendo $\mathbf{0}$ o vetor nulo de dimensão n, $\boldsymbol{I}_n$ a matriz identidade de ordem n e $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ ou

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

Forma matricial

A matriz $\boldsymbol{X}$ tem n linhas (tamanho da amostra) e p colunas (número de variáveis preditoras, covariáveis), de posto p ($n \geq p$), $\boldsymbol{X}$ é chamada de matriz de planejamento ou desenho.

A forma matricial é bastante útil para a implementação computacional, pois ela evita *looping*.

Forma matricial

A matriz $\boldsymbol{X}$ tem n linhas (tamanho da amostra) e p colunas (número de variáveis preditoras, covariáveis), de posto p ($n \geq p$), $\boldsymbol{X}$ é chamada de matriz de planejamento ou desenho.

A forma matricial é bastante útil para a implementação computacional, pois ela evita *looping*.

Normalidade

Para que procedimentos inferenciais possam ser aplicados, nós precisamos assumir que, $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$ e $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, em (1) e (3), respectivamente. Com essa suposição, o MRL passa a ser denominado de **modelo de regressão normal linear** (MNL).

Normalidade

Para que procedimentos inferenciais possam ser aplicados, nós precisamos assumir que, $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$ e $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, em (1) e (3), respectivamente. Com essa suposição, o MRL passa a ser denominado de **modelo de regressão normal linear** (MNL).

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

Normalidade

Para que procedimentos inferenciais possam ser aplicados, nós precisamos assumir que, $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$ e $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, em (1) e (3), respectivamente. Com essa suposição, o MRL passa a ser denominado de **modelo de regressão normal linear** (MNL).

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

Exemplos

Exemplo 1. Conjunto de dados *Westwood Company* (Neter et al., 1983, p. 36), também disponíveis na Tabela 1.

Exemplos

Exemplo 1. Conjunto de dados *Westwood Company* (Neter et al., 1983, p. 36), também disponíveis na Tabela 1.

Tabela 1: Tamanho do lote e horas trabalhadas.

Lote	30	20	60	80	40	50	60	30	70	60
Horas	73	50	128	170	87	108	135	69	148	132

Exemplos

Exemplo 1. Conjunto de dados *Westwood Company* (Neter et al., 1983, p. 36), também disponíveis na Tabela 1.

Tabela 1: Tamanho do lote e horas trabalhadas.

Lote	30	20	60	80	40	50	60	30	70	60
Horas	73	50	128	170	87	108	135	69	148	132

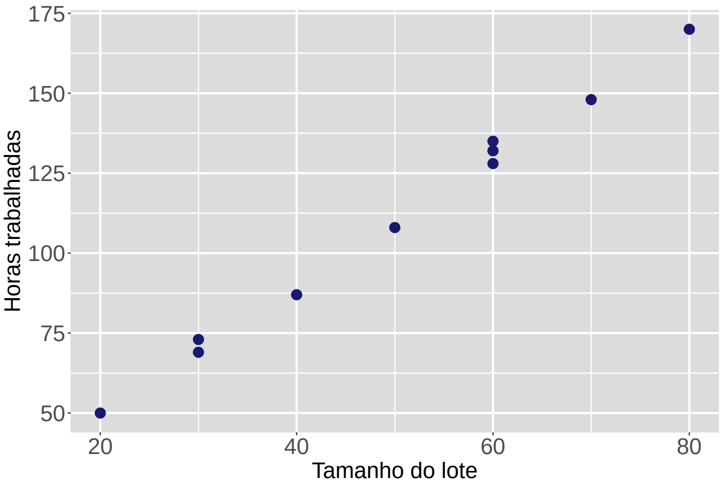

Figura 1: Gráfico de dispersão para os dados *Westwood Company*.

Exemplos

Pela Figura 1, nós poderíamos supor um MRL,

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

Exemplos

Pela Figura 1, nós poderíamos supor um MRL,

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

em que Y_ℓ: horas trabalhadas, $x_{\ell 2}$: tamanho do lote, $\ell = 1, 2, \ldots, 10$.

Exemplos

Pela Figura 1, nós poderíamos supor um MRL,

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

em que Y_ℓ: horas trabalhadas, $x_{\ell 2}$: tamanho do lote, $\ell = 1, 2, \ldots, 10$.

Exemplos

Exemplo 2. Conjunto de dados “propelente de foguetes” (Montgomery et al., 2012, p. 15), também disponíveis na Tabela 2.

Exemplos

Exemplo 2. Conjunto de dados "propelente de foguetes" (Montgomery et al., 2012, p. 15), também disponíveis na Tabela 2.

Exemplo 2. Conjunto de dados “propelente de foguetes” (Montgomery et al., 2012, p. 15), também disponíveis na Tabela 2.

Tabela 2: Propelente de foguetes.

Resistência	Idade	Resistência	Idade
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

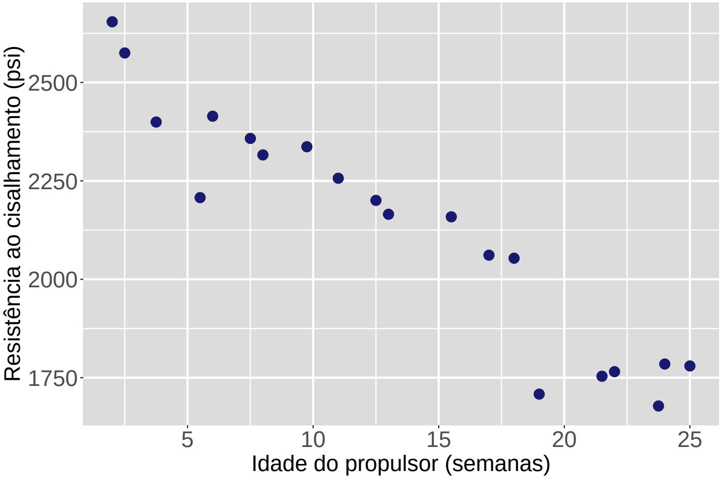

Figura 2: Gráfico de dispersão para os dados de propelente de foguete.

Exemplos

Pela Figura 2, nós poderíamos supor um MRL,

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

Exemplos

Pela Figura 2, nós poderíamos supor um MRL,

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

em que Y_ℓ: resistência ao cisalhamento (psi), $x_{\ell 2}$: idade do propulsor (semanas), $\ell = 1, 2, \ldots, 20$.

Exemplos

Pela Figura 2, nós poderíamos supor um MRL,

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

em que Y_ℓ: resistência ao cisalhamento (psi), $x_{\ell 2}$: idade do propulsor (semanas), $\ell = 1, 2, \ldots, 20$.

Roteiro

Padronização

Em algumas situações, é conveniente mudar a escala das variáveis. Por exemplo, quando:

- há alta correlação entre variáveis preditoras;

Padronização

Em algumas situações, é conveniente mudar a escala das variáveis. Por exemplo, quando:

- há alta correlação entre variáveis preditoras;
- o intervalo de variação das covariáveis é muito diferente.

Padronização

Em algumas situações, é conveniente mudar a escala das variáveis. Por exemplo, quando:

- há alta correlação entre variáveis preditoras;
- o intervalo de variação das covariáveis é muito diferente.

Padronização

A primeira possibilidade é a seguinte:

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

Padronização

A primeira possibilidade é a seguinte:

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

em que $x^0_{\ell m}$ é um valor de referência, $\ell = 1, 2, \ldots, n$ e $m = 1, 2, \ldots, p$.

Comumente, se utiliza $x^0_m = \bar{x}_m$, a média amostral da m-ésima covariável.

Padronização

A primeira possibilidade é a seguinte:

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

em que $x^0_{\ell m}$ é um valor de referência, $\ell = 1, 2, \ldots, n$ e $m = 1, 2, \ldots, p$.

Comumente, se utiliza $x^0_m = \bar{x}_m$, a média amostral da m-ésima covariável.

Escala normal de unidade

Na segunda possibilidade, denominada de **escala normal de unidade**, o procedimento é o seguinte:

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

Escala normal de unidade

Na segunda possibilidade, denominada de **escala normal de unidade**, o procedimento é o seguinte:

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

em que S_Y e s_m são, respectivamente, as raízes quadradas de

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

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

Escala normal de unidade

Na segunda possibilidade, denominada de **escala normal de unidade**, o procedimento é o seguinte:

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

em que S_Y e s_m são, respectivamente, as raízes quadradas de

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

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

Escala normal de unidade

Observação

Todos os regressores transformados e as respostas transformadas têm média amostral zero e variância amostral um.

Escala normal de unidade

Observação

Todos os regressores transformados e as respostas transformadas têm média amostral zero e variância amostral um.

Escala de tamanho unitário

Na terceira possibilidade, denominada de **escala tamanho unitário**, o procedimento é o seguinte:

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

Escala de tamanho unitário

Na terceira possibilidade, denominada de **escala tamanho unitário**, o procedimento é o seguinte:

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

em que S_Y e s_m são, respectivamente, as raízes quadradas de

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

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

Escala de tamanho unitário

Na terceira possibilidade, denominada de **escala tamanho unitário**, o procedimento é o seguinte:

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

em que S_Y e s_m são, respectivamente, as raízes quadradas de

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

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

Escala tamanho unitário

Observação

Cada regressor e a resposta têm média amostral zero e tamanho unitário, ou seja,

Escala tamanho unitário

Observação

Cada regressor e a resposta têm média amostral zero e tamanho unitário, ou seja,

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

Escala tamanho unitário

Observação

Cada regressor e a resposta têm média amostral zero e tamanho unitário, ou seja,

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

Padronização

Observação final

Notem que, em um modelo com intercepto, o procedimento de padronização das variáveis, o removerá do modelo.

Padronização

Observação final

Notem que, em um modelo com intercepto, o procedimento de padronização das variáveis, o removerá do modelo.

Roteiro

1. Introdução
2. Modelo de regressão linear
3. Padronização
4. Covariáveis qualitativas
5. Estudo de simulação
6. Considerações finais
7. Referências bibliográficas

Covariáveis qualitativas

Não é incomum que variáveis preditoras qualitativas (categóricas) sejam utilizadas para explicar o comportamento de uma variável resposta. Porém, o seu uso exige um pouco mais de atenção no tratamento.

Covariáveis qualitativas

Não é incomum que variáveis preditoras qualitativas (categóricas) sejam utilizadas para explicar o comportamento de uma variável resposta. Porém, o seu uso exige um pouco mais de atenção no tratamento.

Covariáveis qualitativas

Situação 1. Suponham o seguinte MRL,

Situação 1. Suponham o seguinte MRL,

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

Covariáveis qualitativas

Situação 1. Suponham o seguinte MRL,

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

em que $x_{\ell 2}$ é uma variável contínua qualquer e $x_{\ell 3}$ é o sexo do indivíduo ℓ, $x_{\ell 3} \in \{\text{feminino}, \text{masculino}\}$, $\ell = 1, 2, \ldots, n$.

Covariáveis qualitativas

Situação 1. Suponham o seguinte MRL,

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

em que $x_{\ell 2}$ é uma variável contínua qualquer e $x_{\ell 3}$ é o sexo do indivíduo ℓ, $x_{\ell 3} \in \{\text{feminino}, \text{masculino}\}$, $\ell = 1, 2, \ldots, n$.

Covariáveis qualitativas

Suponham, também, que

$$x_{\ell 3} = \begin{cases} 1, & \text{se é do sexo feminino} \\ 0, & \text{se é do sexo masculino} \end{cases}, \quad \ell = 1, 2, \ldots, n.$$

Covariáveis qualitativas

Suponham, também, que

$$x_{\ell 3} = \begin{cases} 1, & \text{se é do sexo feminino} \\ 0, & \text{se é do sexo masculino} \end{cases}, \;\; \ell = 1, 2, \ldots, n.$$

A variável $x_{\ell 3}$, escrita da forma acima, é chamada de variável indicadora ou **variável *dummy***.

Covariáveis qualitativas

Suponham, também, que

$$x_{\ell 3} = \begin{cases} 1, & \text{se é do sexo feminino} \\ 0, & \text{se é do sexo masculino} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

A variável $x_{\ell 3}$, escrita da forma acima, é chamada de variável indicadora ou **variável *dummy***. Agora, notem que,

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

Covariáveis qualitativas

Suponham, também, que

$$x_{\ell 3} = \begin{cases} 1, & \text{se é do sexo feminino} \\ 0, & \text{se é do sexo masculino} \end{cases}, \quad \ell = 1, 2, \ldots, n.$$

A variável $x_{\ell 3}$, escrita da forma acima, é chamada de variável indicadora ou **variável *dummy***. Agora, notem que,

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

Isto é, o MRL se converteu em dois MLS.

Covariáveis qualitativas

Suponham, também, que

$$x_{\ell 3} = \begin{cases} 1, & \text{se é do sexo feminino} \\ 0, & \text{se é do sexo masculino} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

A variável $x_{\ell 3}$, escrita da forma acima, é chamada de variável indicadora ou **variável *dummy***. Agora, notem que,

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

Isto é, o MRL se converteu em dois MLS.

Situação 2. Agora, suponham o seguinte MRL,

Covariáveis qualitativas

Situação 2. Agora, suponham o seguinte MRL,

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

Situação 2. Agora, suponham o seguinte MRL,

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

em que $x_{\ell 2}$ é uma variável contínua qualquer e $x_{\ell 3}$ é a escolaridade do indivíduo ℓ, $x_{\ell 3} \in \{\text{fundamental}, \text{médio}, \text{superior}\}$, $\ell = 1, 2, \ldots, n$.

Covariáveis qualitativas

Situação 2. Agora, suponham o seguinte MRL,

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

em que $x_{\ell 2}$ é uma variável contínua qualquer e $x_{\ell 3}$ é a escolaridade do indivíduo ℓ, $x_{\ell 3} \in \{\text{fundamental}, \text{médio}, \text{superior}\}$, $\ell = 1, 2, \ldots, n$.

Covariáveis qualitativas

Para essa situação, é necessário criar duas variáveis *dummies*:

$$x_{\ell 3} = \begin{cases} 1, & \text{se tem ensino fundamental} \\ 0, & \text{caso contrário} \end{cases},$$

$$x_{\ell 4} = \begin{cases} 1, & \text{se tem ensino médio} \\ 0, & \text{caso contrário} \end{cases}, \ \ell = 1, 2, \ldots, n.$$

Covariáveis qualitativas

Para essa situação, é necessário criar duas variáveis *dummies*:

$$x_{\ell 3} = \begin{cases} 1, & \text{se tem ensino fundamental} \\ 0, & \text{caso contrário} \end{cases},$$

$$x_{\ell 4} = \begin{cases} 1, & \text{se tem ensino médio} \\ 0, & \text{caso contrário} \end{cases}, \;\ell = 1, 2, \ldots, n.$$

Agora, notem que,

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

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

Agora, notem que,

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

$\ell = 1, 2, \ldots, n$. Isto é, o MRL se converteu em três MLS. O mais importante é que, uma variável qualitativa muda a matriz de planejamento.

Covariáveis qualitativas

Agora, notem que,

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

$\ell = 1, 2, \ldots, n$. Isto é, o MRL se converteu em três MLS. O mais importante é que, uma variável qualitativa muda a matriz de planejamento.

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

Covariáveis qualitativas

Se uma variável preditora possui k categorias, serão necessárias $k-1$ variáveis *dummies*.

Exemplo 3. Conjunto de dados de Walker (1962). O autor estudou os sons de grilos machos nos EUA, onde ele mensurou o número de pulsações por segundo de 31 grilos sob diferentes temperaturas, também disponíveis na Tabela 3.

Tabela 3: Número de pulsações dos grilos.

Esp.	Temp.	Pul.	Esp.	Temp.	Pul.	Esp.	Temp.	Pul.
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			

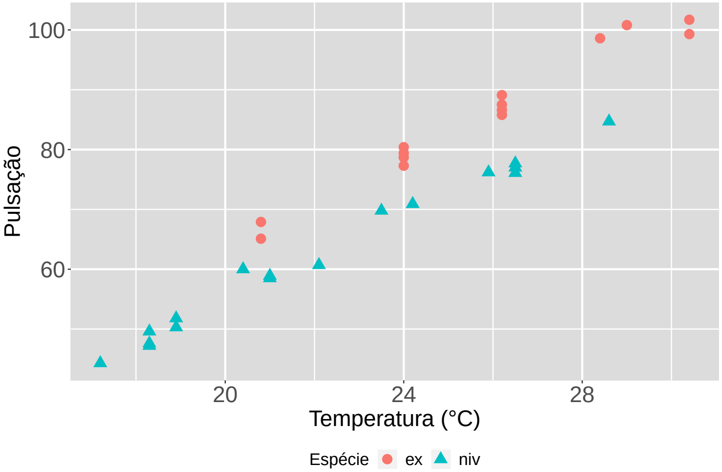

Figura 3: Gráfico de dispersão número de pulsações pela temperatura.

Exemplos

Pela Figura 3, nós poderíamos supor um MRL,

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

Exemplos

Pela Figura 3, nós poderíamos supor um MRL,

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

em que Y_ℓ: número de pulsações, $x_{\ell 2}$: temperatura, $x_{\ell 3}$ a espécie do grilo, $x_{\ell 3} \in \{\textit{Oecanthus exclamationis (ex)}, \textit{Oecanthus niveus (niv)}\}$, $\ell = 1, 2, \ldots, 31$.

Exemplos

Pela Figura 3, nós poderíamos supor um MRL,

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

em que Y_ℓ: número de pulsações, $x_{\ell 2}$: temperatura, $x_{\ell 3}$ a espécie do grilo, $x_{\ell 3} \in \{\textit{Oecanthus exclamationis (ex)}, \textit{Oecanthus niveus (niv)}\}$, $\ell = 1, 2, \ldots, 31$.

Roteiro

Estudo de simulação

Suponham o seguinte MNL,

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

Estudo de simulação

Suponham o seguinte MNL,

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

em que $x_{\ell 2}$ é quantidade conhecida e $\varepsilon_\ell \sim \mathcal{N}(0, 4)$, $\ell = 1, 2, \ldots, n$.

Estudo de simulação

Suponham o seguinte MNL,

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

em que $x_{\ell 2}$ é quantidade conhecida e $\varepsilon_\ell \sim \mathcal{N}(0, 4)$, $\ell = 1, 2, \ldots, n$. Notem que, $Y_\ell \sim \mathcal{N}(1+x_{\ell 2}, 4)$ e nós vamos assumir que $x_{\ell 2} \sim$ uniforme no intervalo (0,1).

Estudo de simulação

Suponham o seguinte MNL,

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

em que $x_{\ell 2}$ é quantidade conhecida e $\varepsilon_\ell \sim \mathcal{N}(0,4)$, $\ell = 1, 2, \ldots, n$. Notem que, $Y_\ell \sim \mathcal{N}(1+x_{\ell 2}, 4)$ e nós vamos assumir que $x_{\ell 2} \sim$ uniforme no intervalo (0,1).

Nós faremos um estudo de Monte Carlo, com 5.000 réplicas, onde avaliaremos $\text{Cov}(Y, x_2)$, para $n \in \{10, 20, 40, 80, 160\}$.

Estudo de simulação

Suponham o seguinte MNL,

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

em que $x_{\ell 2}$ é quantidade conhecida e $\varepsilon_\ell \sim \mathcal{N}(0,4)$, $\ell = 1, 2, \ldots, n$. Notem que, $Y_\ell \sim \mathcal{N}(1+x_{\ell 2}, 4)$ e nós vamos assumir que $x_{\ell 2} \sim$ uniforme no intervalo (0,1).

Nós faremos um estudo de Monte Carlo, com 5.000 réplicas, onde avaliaremos $\text{Cov}(Y, x_2)$, para $n \in \{10, 20, 40, 80, 160\}$.

Estudo de simulação

Isto é,

1. Nós fixamos os parâmetros, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ e $n = 10$;

Estudo de simulação

Isto é,

1. Nós fixamos os parâmetros, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ e $n = 10$;
2. Nós geramos 10 valores de $\boldsymbol{x}_2$ a partir de uma uniforme(0, 1);

Estudo de simulação

Isto é,

1. Nós fixamos os parâmetros, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ e $n = 10$;
2. Nós geramos 10 valores de $\boldsymbol{x}_2$ a partir de uma uniforme(0, 1);
3. Para cada replicação, nós geramos 10 valores de $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;

Estudo de simulação

Isto é,

1. Nós fixamos os parâmetros, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ e $n = 10$;
2. Nós geramos 10 valores de $\boldsymbol{x}_2$ a partir de uma uniforme(0, 1);
3. Para cada replicação, nós geramos 10 valores de $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;
4. Também para cada replicação, nós calculamos $\text{Cov}(Y, x_2)$;

Estudo de simulação

Isto é,

1. Nós fixamos os parâmetros, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ e $n = 10$;
2. Nós geramos 10 valores de $\boldsymbol{x}_2$ a partir de uma uniforme(0, 1);
3. Para cada replicação, nós geramos 10 valores de $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;
4. Também para cada replicação, nós calculamos $\text{Cov}(Y, x_2)$;
5. Com as 5.000 réplicas, nós calculamos a média das 5.000 covariâncias.

Estudo de simulação

Isto é,

1. Nós fixamos os parâmetros, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ e $n = 10$;
2. Nós geramos 10 valores de $\boldsymbol{x}_2$ a partir de uma uniforme(0, 1);
3. Para cada replicação, nós geramos 10 valores de $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;
4. Também para cada replicação, nós calculamos Cov(Y, x_2);
5. Com as 5.000 réplicas, nós calculamos a média das 5.000 covariâncias.

O experimento acima foi repetido mais 4 vezes. Alterando somente o tamanho de amostra para $20, 40, 80, 160$.

Estudo de simulação

Isto é,

1. Nós fixamos os parâmetros, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma^2 = 4$ e $n = 10$;
2. Nós geramos 10 valores de $\boldsymbol{x}_2$ a partir de uma uniforme(0, 1);
3. Para cada replicação, nós geramos 10 valores de $Y_\ell \sim \mathcal{N}(1 + x_{\ell 2}, 4)$;
4. Também para cada replicação, nós calculamos Cov(Y, x_2);
5. Com as 5.000 réplicas, nós calculamos a média das 5.000 covariâncias.

O experimento acima foi repetido mais 4 vezes. Alterando somente o tamanho de amostra para $20, 40, 80, 160$.

Estudo de simulação

Tabela 4: Resultados.

n	Covariância	DP
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

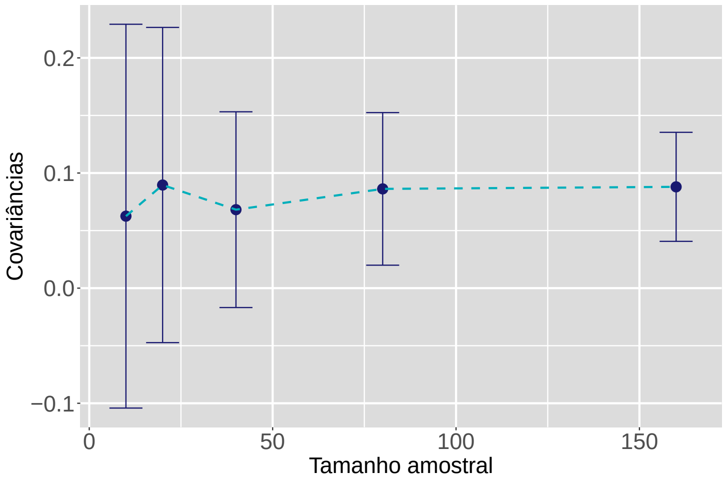

Figura 4: Covariâncias em relação ao tamanho amostral.

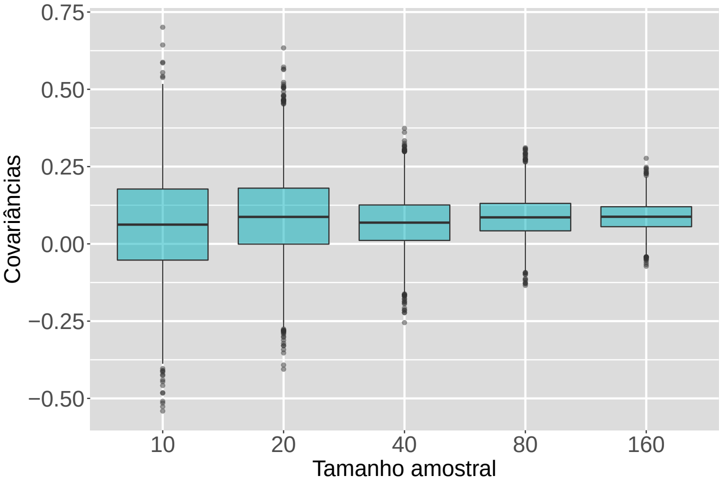

Figura 5: Boxplot das covariâncias por tamanho amostral.

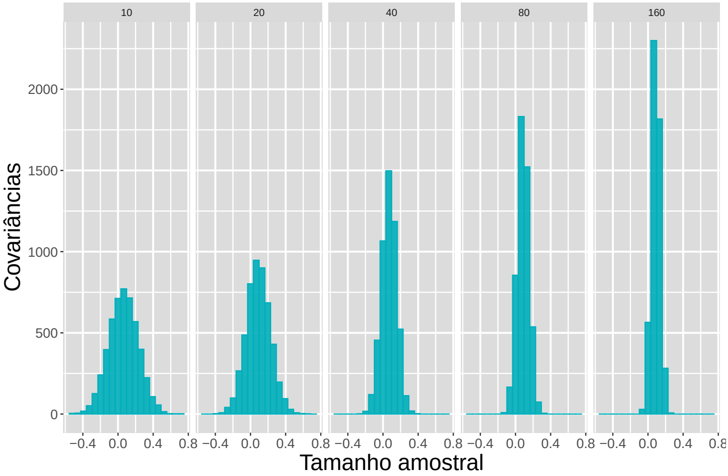

Figura 6: Histograma das covariâncias por tamanho amostral.

Estudo de simulação

Comentários

Pelas Figuras 4 a 6, nós podemos perceber que as estimativas das covariâncias entre Y e x_2 parecem convergir para um **certo valor**.

Estudo de simulação

Comentários

Pelas Figuras 4 a 6, nós podemos perceber que as estimativas das covariâncias entre Y e x_2 parecem convergir para um **certo valor**. Como esperado, a medida que o tamanho de amostra aumenta, a variabilidade das estimativas das covariâncias diminui.

Comentários

Pelas Figuras 4 a 6, nós podemos perceber que as estimativas das covariâncias entre Y e x_2 parecem convergir para um **certo valor**. Como esperado, a medida que o tamanho de amostra aumenta, a variabilidade das estimativas das covariâncias diminui.

Roteiro

Considerações finais

Para um melhor entendimento dos conceitos de modelos de regressão linear, ver os Capítulos 1 de Kutner et al. (2005) e Montgomery et al. (2012).

Roteiro

Referências bibliográficas I

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

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

Montgomery, D. C., Peck, E. A. e Vining, G. G. (2012), *Introduction to linear regression analysis*, 5th edn, Wiley, New York.

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

Referências bibliográficas II

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.

Obrigado!

✉ tiago.magalhaes@ufjf.br

ufjf.br/tiago_magalhaes

Departamento de Estatística, Sala 319