Transformações

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 06 de maio de 2024

Roteiro

Roteiro

Modelo de regressão linear

Suponham que $Y_1, Y_2, \ldots, Y_n$ tais que

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \ \ell = 1, 2, \ldots, n, \tag{1}$$

Modelo de regressão linear

Suponham que $Y_1, Y_2, \ldots, Y_n$ tais que

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \ \ell = 1, 2, \ldots, n, \tag{1}$$

em que $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ é conhecido, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ é um vetor de parâmetros desconhecidos a serem estimados, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2, também desconhecida, a ser estimada.

Modelo de regressão linear

Suponham que $Y_1, Y_2, \ldots, Y_n$ tais que

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \ \ell = 1, 2, \ldots, n, \tag{1}$$

em que $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ é conhecido, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ é um vetor de parâmetros desconhecidos a serem estimados, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2, também desconhecida, a ser estimada.

Forma matricial

A Equação (1) pode ser escrita de forma matricial, da seguinte forma:

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

Forma matricial

A Equação (1) pode ser escrita de forma matricial, da seguinte forma:

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

em que $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ é a matriz de planejamento e $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, com

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ e } \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n.$$

Forma matricial

A Equação (1) pode ser escrita de forma matricial, da seguinte forma:

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

em que $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ é a matriz de planejamento e $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, com

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ e } \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n.$$

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;
  - variância constante;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;
  - variância constante;
  - e são não correlacionados.

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;
  - variância constante;
  - e são não correlacionados.

Roteiro

Transformações estabilizadoras da variância

A suposição de variância constante costuma ser violada quando ela está relacionada com a média.

Um modelo com variância não constante costuma estimar os coeficientes de regressão com erros padrão elevados.

Transformações estabilizadoras da variância

A suposição de variância constante costuma ser violada quando ela está relacionada com a média.

Um modelo com variância não constante costuma estimar os coeficientes de regressão com erros padrão elevados.

Transformações estabilizadoras da variância

Transformações na variável resposta podem eliminar este problema.

O efeito da transformação depende da quantidade de curvatura que ela induz.

Transformações estabilizadoras da variância

Transformações na variável resposta podem eliminar este problema.

O efeito da transformação depende da quantidade de curvatura que ela induz.

Transformações estabilizadoras da variância

Tabela 1: Transformações úteis.

Relação entre σ^2 e $\mathbb{E}(Y)$	Transformação	Observação
$\sigma^2 \propto$ constante	$Y^* = Y$	Sem transformação
$\sigma^2 \propto \mathbb{E}(Y)$	$Y^* = \sqrt{Y}$	Dados Poisson
$\sigma^2 \propto \mathbb{E}(Y)[1 - \mathbb{E}(Y)]$	$Y^* =$ arcoseno(Y)	Dados binomial
$\sigma^2 \propto [\mathbb{E}(Y)]^2$	$Y^* = \log(Y)$	
$\sigma^2 \propto [\mathbb{E}(Y)]^3$	$Y^* = Y^{-1/2}$	
$\sigma^2 \propto [\mathbb{E}(Y)]^4$	$Y^* = Y^{-1}$	

Roteiro

Transformação Box-Cox

Uma classe de transformações úteis para corrigir não normalidade e/ou variância não constante.

Proposta por Box e Cox (1964), é uma classe de transformações dadas por

$$Y^* = Y(\lambda),$$

em que λ é um parâmetro adicional a ser estimado.

Transformação Box-Cox

Uma classe de transformações úteis para corrigir não normalidade e/ou variância não constante.

Proposta por Box e Cox (1964), é uma classe de transformações dadas por

$$Y^* = Y(\lambda),$$

em que λ é um parâmetro adicional a ser estimado.

Transformação Box-Cox

O procedimento é o seguinte:

$$Y^* = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \text{se } \lambda \neq 0; \\ \log Y, & \text{se } \lambda = 0. \end{cases} \tag{3}$$

Transformação Box-Cox

- Assim como para os demais parâmetros,

Transformação Box-Cox

- Assim como para os demais parâmetros, λ pode ser estimado por máxima verossimilhança.

Transformação Box-Cox

- Assim como para os demais parâmetros, λ pode ser estimado por máxima verossimilhança.
- O estimador de λ não tem forma fechada

Transformação Box-Cox

- Assim como para os demais parâmetros, λ pode ser estimado por máxima verossimilhança.
- O estimador de λ não tem forma fechada e assim métodos numéricos são necessários para se encontrar a estimativa de máxima verossimilhança para λ.

Transformação Box-Cox

- Assim como para os demais parâmetros, λ pode ser estimado por máxima verossimilhança.
- O estimador de λ não tem forma fechada e assim métodos numéricos são necessários para se encontrar a estimativa de máxima verossimilhança para λ.

Transformação Box-Cox

Existem versões alternativas para (3), entre elas:

$$Y^* = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda \dot{Y}^{\lambda-1}}, & \text{se } \lambda \neq 0; \\ \dot{Y} \log Y, & \text{se } \lambda = 0; \end{cases}$$

em que $\dot{Y}$ é a média geométrica.

Transformação Box-Cox

Existem versões alternativas para (3), entre elas:

$$Y^* = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda \dot{Y}^{\lambda-1}}, & \text{se } \lambda \neq 0; \\ \\ \dot{Y} \log Y, & \text{se } \lambda = 0; \end{cases}$$

em que $\dot{Y}$ é a média geométrica.

Roteiro

Mínimos quadrados generalizados

Um modelo de regressão linear com variância não constante também pode ser ajustado através do método de mínimos quadrados generalizados.

Neste método de estimação, os desvios entre o valor observado e esperado da variável Y_ℓ é multiplicada por um peso w_ℓ,

Mínimos quadrados generalizados

Um modelo de regressão linear com variância não constante também pode ser ajustado através do método de mínimos quadrados generalizados.

Neste método de estimação, os desvios entre o valor observado e esperado da variável Y_ℓ é multiplicada por um peso w_ℓ, escolhido de forma que ele seja inversamente proporcional a variância de Y_ℓ, $\ell = 1, 2, \ldots, n$.

Mínimos quadrados generalizados

Um modelo de regressão linear com variância não constante também pode ser ajustado através do método de mínimos quadrados generalizados.

Neste método de estimação, os desvios entre o valor observado e esperado da variável Y_ℓ é multiplicada por um peso w_ℓ, escolhido de forma que ele seja inversamente proporcional a variância de Y_ℓ, $\ell = 1, 2, \ldots, n$.

Mínimos quadrados generalizados

Neste método, nós estamos assumindo um modelo de regressão linear, com forma matricial dada em (2),

Mínimos quadrados generalizados

Neste método, nós estamos assumindo um modelo de regressão linear, com forma matricial dada em (2), com

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ e } \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{V}.$$

Mínimos quadrados generalizados

Neste método, nós estamos assumindo um modelo de regressão linear, com forma matricial dada em (2), com

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ e } \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{V}.$$

O **estimador de mínimos quadrados generalizados** de β é dado por:

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

Mínimos quadrados generalizados

Neste método, nós estamos assumindo um modelo de regressão linear, com forma matricial dada em (2), com

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ e } \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{V}.$$

O **estimador de mínimos quadrados generalizados** de $\boldsymbol{\beta}$ é dado por:

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

Mínimos quadrados generalizados

Quando os erros têm variância não constante, mas são não correlacionados,

$$\sigma^2 \boldsymbol{V} = \begin{bmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n \end{bmatrix},$$

Mínimos quadrados generalizados

Quando os erros têm variância não constante, mas são não correlacionados,

$$\sigma^2 \boldsymbol{V} = \begin{bmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n \end{bmatrix},$$

o procedimento é usualmente chamado de método dos mínimos quadrados ponderados.

Mínimos quadrados generalizados

Quando os erros têm variância não constante, mas são não correlacionados,

$$\sigma^2 \boldsymbol{V} = \begin{bmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n \end{bmatrix},$$

o procedimento é usualmente chamado de método dos mínimos quadrados ponderados. E o estimador fica: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{Y}$.

Mínimos quadrados generalizados

Quando os erros têm variância não constante, mas são não correlacionados,

$$\sigma^2 \boldsymbol{V} = \begin{bmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n \end{bmatrix},$$

o procedimento é usualmente chamado de método dos mínimos quadrados ponderados. E o estimador fica: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{Y}$.

Mínimos quadrados generalizados

Considerações

A grande de dificuldade deste procedimento é que a matriz $\boldsymbol{V}$ precisa ser conhecida.

Mínimos quadrados generalizados

Considerações

A grande de dificuldade deste procedimento é que a matriz $\boldsymbol{V}$ precisa ser conhecida. Caso contrário, um "chute" inicial precisará ser dado

Mínimos quadrados generalizados

Considerações

A grande de dificuldade deste procedimento é que a matriz $\boldsymbol{V}$ precisa ser conhecida. Caso contrário, um "chute" inicial precisará ser dado e reestimações precisarão ser feitas.

Mínimos quadrados generalizados

Considerações

A grande de dificuldade deste procedimento é que a matriz $\boldsymbol{V}$ precisa ser conhecida. Caso contrário, um "chute" inicial precisará ser dado e reestimações precisarão ser feitas.

Roteiro

Aplicação

(Montgomery et al., 2021, p. 180) Uma empresa de energia investiga a relação entre a demanda no horário de pico (Y, em kW) e total de energia (x_2, em kWh) utilizada durante o mês de agosto,

Aplicação

(Montgomery et al., 2021, p. 180) Uma empresa de energia investiga a relação entre a demanda no horário de pico (Y, em kW) e total de energia (x_2, em kWh) utilizada durante o mês de agosto, com o gráfico de dispersão apresentado na Figura 1.

Aplicação

(Montgomery et al., 2021, p. 180) Uma empresa de energia investiga a relação entre a demanda no horário de pico (Y, em kW) e total de energia (x_2, em kWh) utilizada durante o mês de agosto, com o gráfico de dispersão apresentado na Figura 1. Após o ajuste dos dados, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = -0{,}831 + 0{,}004 x_{\ell 2},$$

$\ell = 1, 2, \ldots, 53.$

Aplicação

(Montgomery et al., 2021, p. 180) Uma empresa de energia investiga a relação entre a demanda no horário de pico (Y, em kW) e total de energia (x_2, em kWh) utilizada durante o mês de agosto, com o gráfico de dispersão apresentado na Figura 1. Após o ajuste dos dados, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = -0{,}831 + 0{,}004 x_{\ell 2},$$

$\ell = 1, 2, \ldots, 53$.

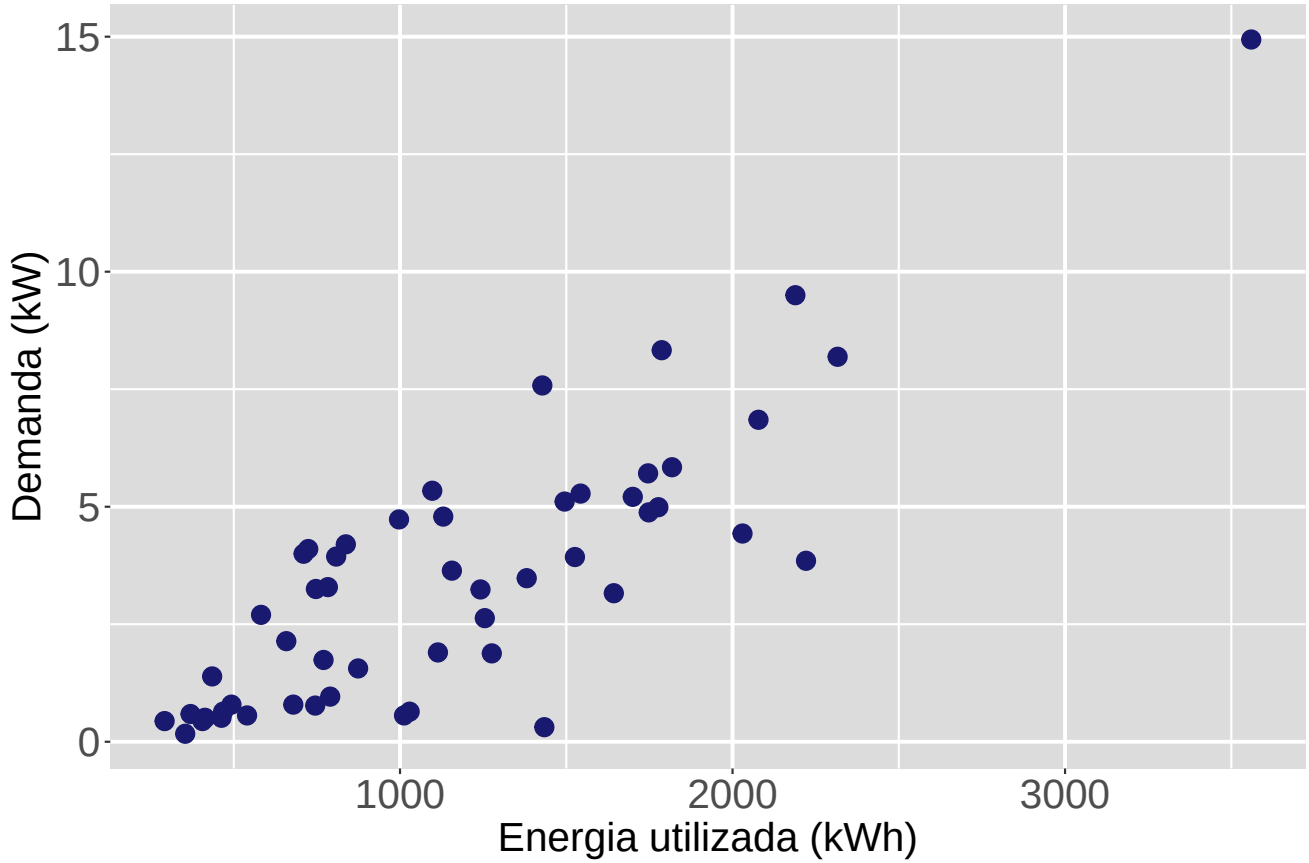


Figura 1: Gráfico de dispersão.

Exemplo

Nós temos também que:

Tabela 2: Estimativas do parâmetros.

Parâmetro	Estimativa	EP	t_c
β_1	-0,831	0,442	-1,882
β_2	0,004	0,001	11,030

Região crítica, para $\alpha = 5\%$: $|t_c| > 2{,}008$, com QMRes $= 2{,}488$.

Exemplo

Nós temos também que:

Tabela 2: Estimativas do parâmetros.

Parâmetro	Estimativa	EP	t_c
β_1	-0,831	0,442	-1,882
β_2	0,004	0,001	11,030

Região crítica, para $\alpha = 5\%$: $|t_c| > 2{,}008$, com QMRes $= 2{,}488$.

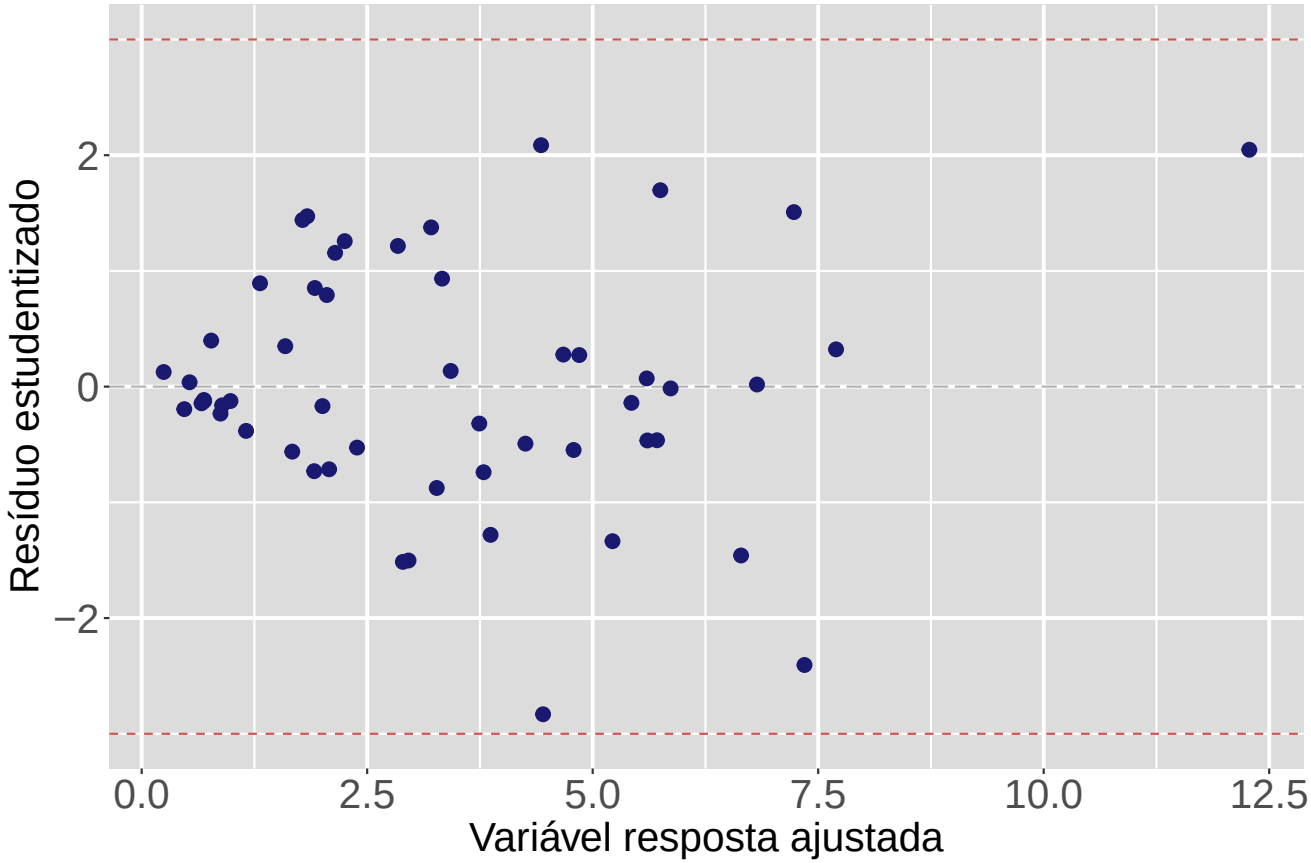


Figura 2: Gráfico de resíduos.

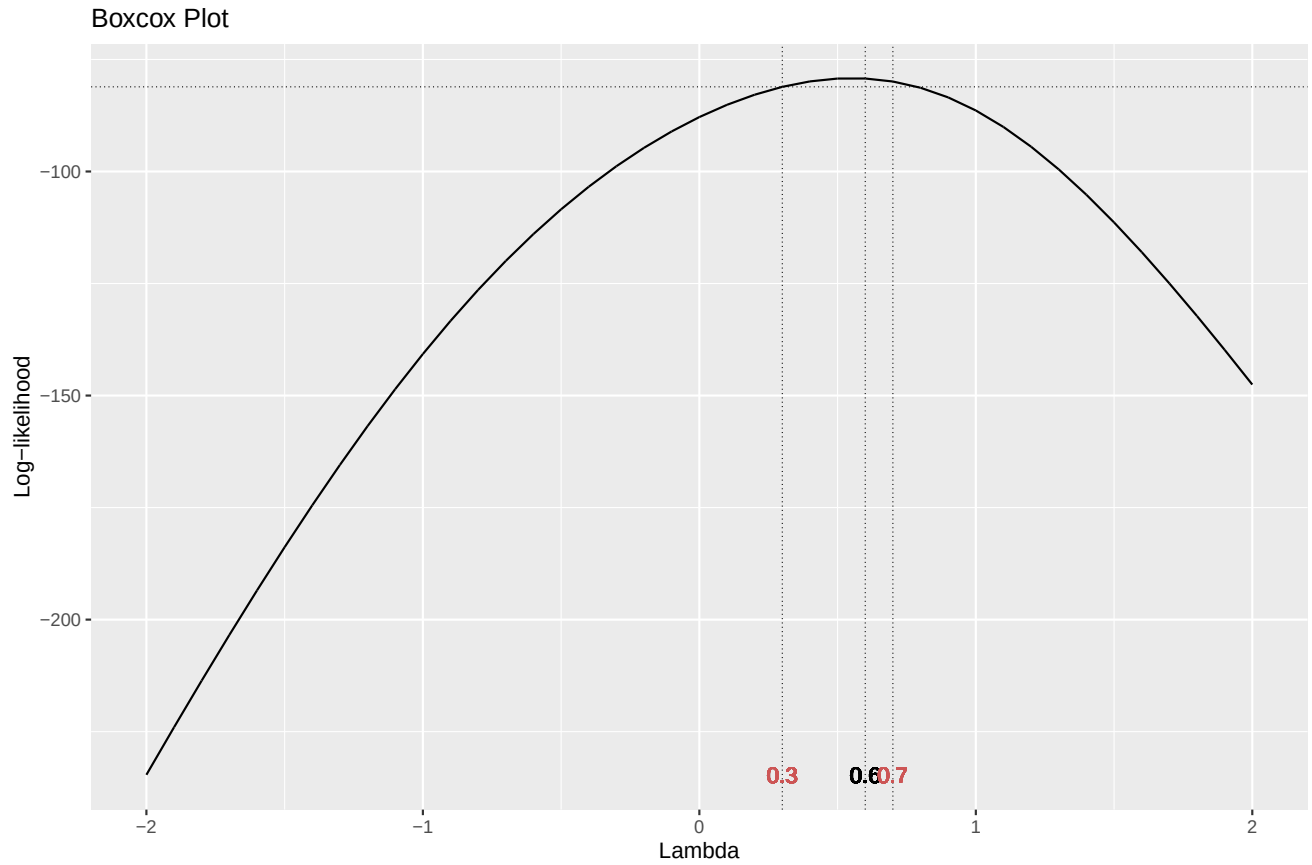


Figura 3: Gráfico Box-Cox, $\lambda = 0{,}5454$.

Aplicação

O método Box-Cox indicou utilizar (3), com $\lambda = 0{,}5454$. Mas, para fins didáticos, nós utilizaremos a seguinte transformação: $Y_\ell^* = \sqrt{Y_\ell}$.

Aplicação

O método Box-Cox indicou utilizar (3), com $\lambda = 0{,}5454$. Mas, para fins didáticos, nós utilizaremos a seguinte transformação: $Y_\ell^* = \sqrt{Y_\ell}$. Após o ajuste dos dados, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell^* = 0{,}582 + 0{,}001 x_{\ell 2},$$

$\ell = 1, 2, \ldots, 53$.

Aplicação

O método Box-Cox indicou utilizar (3), com $\lambda = 0{,}5454$. Mas, para fins didáticos, nós utilizaremos a seguinte transformação: $Y_\ell^* = \sqrt{Y_\ell}$. Após o ajuste dos dados, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell^* = 0{,}582 + 0{,}001 x_{\ell 2},$$

$\ell = 1, 2, \ldots, 53$.

Exemplo

Nós temos também que:

Tabela 3: Estimativas do parâmetros.

Parâmetro	Estimativa	EP	t_c
β_1	0,582	0,130	4,481
β_2	0,001	0,001	9,699

Região crítica, para $\alpha = 5\%$: $|t_c| > 2{,}008$, com QMRes $= 0{,}215$.

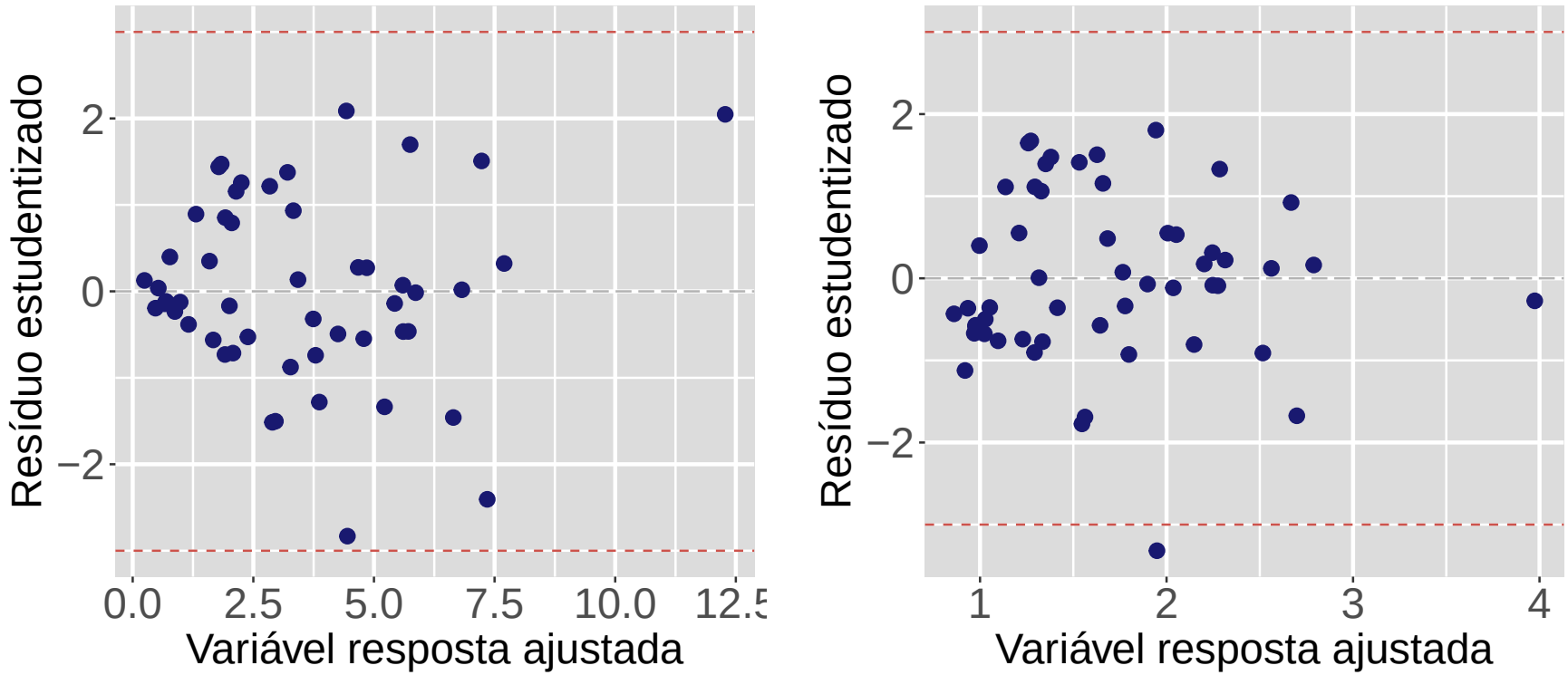
Exemplo

Nós temos também que:

Tabela 3: Estimativas do parâmetros.

Parâmetro	Estimativa	EP	t_c
β_1	0,582	0,130	4,481
β_2	0,001	0,001	9,699

Região crítica, para $\alpha = 5\%$: $|t_c| > 2{,}008$, com QMRes $= 0{,}215$.



(a) Dados originais.

(b) Dados transformados.

Figura 4: Gráficos de resíduos estudentizados.

Roteiro

Referências bibliográficas I

Box, G. E. P. e Cox, D. R. (1964), 'An analysis of transformations (with discussion)', *Journal of the Royal Statistical Society. Series B (Methodological)* **26**(2), 211–252.

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

Obrigado!

✉ tiago.magalhaes@ufjf.br

ufjf.br/tiago_magalhaes

Departamento de Estatística, Sala 319