

# Análise de diagnóstico

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 18 de maio de 2026



# Roteiro

- 1 Introdução
- 2 Pontos de alavanca
- 3 Pontos influentes
- 4 Aplicação
- 5 Referências bibliográficas



# Roteiro

- 1 Introdução
- 2 Pontos de alavanca
- 3 Pontos influentes
- 4 Aplicação
- 5 Referências bibliográficas



# Modelo de regressão linear

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

$$Y_\ell = \mathbf{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \ell = 1, 2, \dots, n, \quad (1)$$



# Modelo de regressão linear

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

$$Y_\ell = \mathbf{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \ell = 1, 2, \dots, n, \quad (1)$$

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



# Modelo de regressão linear

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

$$Y_\ell = \mathbf{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \ell = 1, 2, \dots, n, \quad (1)$$

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



# Forma matricial

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

$$Y = X\beta + \varepsilon, \quad (2)$$



# Forma matricial

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

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

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



# Forma matricial

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2)$$

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

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



# 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

- 1 Introdução
- 2 Pontos de alavanca**
- 3 Pontos influentes
- 4 Aplicação
- 5 Referências bibliográficas



# Pontos de alavanca

Seja a matriz chapéu (*hat matrix*)  $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ , com o seu  $\ell$ -ésimo elemento da diagonal é dado por

$$h_{\ell\ell} = \mathbf{x}_\ell (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_\ell^\top.$$



# Pontos de alavanca

Seja a matriz chapéu (*hat matrix*)  $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ , com o seu  $\ell$ -ésimo elemento da diagonal é dado por

$$h_{\ell\ell} = \mathbf{x}_\ell (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_\ell^\top.$$

$h_{\ell\ell}$  é uma medida de distância entre  $\ell$ -ésima observação e o espaço  $\mathbf{x}$ .



# Pontos de alavanca

Seja a matriz chapéu (*hat matrix*)  $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ , com o seu  $\ell$ -ésimo elemento da diagonal é dado por

$$h_{\ell\ell} = \mathbf{x}_\ell (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_\ell^\top.$$

$h_{\ell\ell}$  é uma medida de distância entre  $\ell$ -ésima observação e o espaço  $\mathbf{x}$ .



# Pontos de alavanca

Observações:

- O tamanho médio da diagonal de  $H$  é  $p/n$ ;



# Pontos de alavanca

Observações:

- O tamanho médio da diagonal de  $\mathbf{H}$  é  $p/n$ ;
- Um ponto é considerado de alavanca, quando  $h_{\ell\ell} > 2p/n$ ;



# Pontos de alavanca

Observações:

- O tamanho médio da diagonal de  $\mathbf{H}$  é  $p/n$ ;
- Um ponto é considerado de alavanca, quando  $h_{\ell\ell} > 2p/n$ ;
- Uma observação com grande  $h_{\ell\ell}$  e um grande resíduo, provavelmente, será um ponto influente.



# Pontos de alavanca

Observações:

- O tamanho médio da diagonal de  $\mathbf{H}$  é  $p/n$ ;
- Um ponto é considerado de alavanca, quando  $h_{\ell\ell} > 2p/n$ ;
- Uma observação com grande  $h_{\ell\ell}$  e um grande resíduo, provavelmente, será um ponto influente.



# Roteiro

- 1 Introdução
- 2 Pontos de alavanca
- 3 Pontos influentes**
- 4 Aplicação
- 5 Referências bibliográficas



# Pontos influentes

As medidas de influência são aquelas que medem o efeito da exclusão da  $l$ -ésima observação.



# Pontos influentes

- Distância de Cook, mede o efeito em  $\hat{\beta}$ ;

# Pontos influentes

- Distância de Cook, mede o efeito em  $\hat{\beta}$ ;
- Dfbetas, mede o efeito em  $\hat{\beta}_m$ ,  $m = 1, 2, \dots, p$ ;



# Pontos influentes

- Distância de Cook, mede o efeito em  $\hat{\beta}$ ;
- Dfbetas, mede o efeito em  $\hat{\beta}_m$ ,  $m = 1, 2, \dots, p$ ;
- Dffits, mede o efeito em  $\hat{Y}_\ell$ ,  $\ell = 1, 2, \dots, n$ ;



# Pontos influentes

- Distância de Cook, mede o efeito em  $\hat{\beta}$ ;
- Dfbetas, mede o efeito em  $\hat{\beta}_m$ ,  $m = 1, 2, \dots, p$ ;
- Dffits, mede o efeito em  $\hat{Y}_\ell$ ,  $\ell = 1, 2, \dots, n$ ;
- Covratio, mede o efeito na matriz de covariâncias.

# Pontos influentes

- Distância de Cook, mede o efeito em  $\hat{\beta}$ ;
- Dfbetas, mede o efeito em  $\hat{\beta}_m$ ,  $m = 1, 2, \dots, p$ ;
- Dffits, mede o efeito em  $\hat{Y}_\ell$ ,  $\ell = 1, 2, \dots, n$ ;
- Covratio, mede o efeito na matriz de covariâncias.

# Pontos influentes

A **distância de Cook** (Cook, 1977, 1979) serve para analisar o quão bem o modelo ajusta a observação  $Y_\ell$  e quão a observação está distante do conjunto de dados.



# Pontos influentes

A **distância de Cook** (Cook, 1977, 1979) serve para analisar o quão bem o modelo ajusta a observação  $Y_\ell$  e quão a observação está distante do conjunto de dados. Esta medida é calculada da seguinte forma:

$$D_\ell = \frac{(\hat{\beta}_{(\ell)} - \hat{\beta})^\top (\mathbf{X}^\top \mathbf{X})(\hat{\beta}_{(\ell)} - \hat{\beta})}{p \text{QMRes}} = \frac{r_\ell^2}{p} \frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$



# Pontos influentes

A **distância de Cook** (Cook, 1977, 1979) serve para analisar o quão bem o modelo ajusta a observação  $Y_\ell$  e quão a observação está distante do conjunto de dados. Esta medida é calculada da seguinte forma:

$$D_\ell = \frac{(\hat{\beta}_{(\ell)} - \hat{\beta})^\top (\mathbf{X}^\top \mathbf{X})(\hat{\beta}_{(\ell)} - \hat{\beta})}{p \text{QMRes}} = \frac{r_\ell^2}{p} \frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

Um valor será considerado influente se  $D_\ell > 1$ .

# Pontos influentes

A **distância de Cook** (Cook, 1977, 1979) serve para analisar o quão bem o modelo ajusta a observação  $Y_\ell$  e quão a observação está distante do conjunto de dados. Esta medida é calculada da seguinte forma:

$$D_\ell = \frac{(\hat{\beta}_{(\ell)} - \hat{\beta})^\top (\mathbf{X}^\top \mathbf{X})(\hat{\beta}_{(\ell)} - \hat{\beta})}{p \text{QMRes}} = \frac{r_\ell^2}{p} \frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

Um valor será considerado influente se  $D_\ell > 1$ . A distância de Cook pode ser utilizada para verificar a influência conjunta de observações.



# Pontos influentes

A **distância de Cook** (Cook, 1977, 1979) serve para analisar o quão bem o modelo ajusta a observação  $Y_\ell$  e quão a observação está distante do conjunto de dados. Esta medida é calculada da seguinte forma:

$$D_\ell = \frac{(\hat{\beta}_{(\ell)} - \hat{\beta})^\top (\mathbf{X}^\top \mathbf{X})(\hat{\beta}_{(\ell)} - \hat{\beta})}{p \text{QMRes}} = \frac{r_\ell^2}{p} \frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

Um valor será considerado influente se  $D_\ell > 1$ . A distância de Cook pode ser utilizada para verificar a influência conjunta de observações.



# Pontos influentes

O **Dfbetas** (Belsley et al., 1980) mede o quanto o coeficiente de regressão se altera, em unidades de desvio padrão, quando a  $\ell$ -ésima observação é removida. Esta medida é calculada da seguinte forma:

$$\text{Dfbetas}_{m,\ell} = \frac{\hat{\beta}_m - \hat{\beta}_{m(\ell)}}{\sqrt{S_{(\ell)}^2 c_{mm}}},$$

## Pontos influentes

O **Dfbetas** (Belsley et al., 1980) mede o quanto o coeficiente de regressão se altera, em unidades de desvio padrão, quando a  $\ell$ -ésima observação é removida. Esta medida é calculada da seguinte forma:

$$\text{Dfbetas}_{m,\ell} = \frac{\hat{\beta}_m - \hat{\beta}_{m(\ell)}}{\sqrt{S_{(\ell)}^2 c_{mm}}},$$

em que  $\hat{\beta}_{m(\ell)}$  é o  $m$ -ésimo coeficiente estimado, quando a  $\ell$ -ésima observação é removida.



# Pontos influentes

O **Dfbetas** (Belsley et al., 1980) mede o quanto o coeficiente de regressão se altera, em unidades de desvio padrão, quando a  $\ell$ -ésima observação é removida. Esta medida é calculada da seguinte forma:

$$\text{Dfbetas}_{m,\ell} = \frac{\hat{\beta}_m - \hat{\beta}_{m(\ell)}}{\sqrt{S_{(\ell)}^2 c_{mm}}},$$

em que  $\hat{\beta}_{m(\ell)}$  é o  $m$ -ésimo coeficiente estimado, quando a  $\ell$ -ésima observação é removida. Um valor será considerado influente se  $\text{Dfbetas}_{m,\ell} > 2/\sqrt{n}$ .



# Pontos influentes

O **Dfbetas** (Belsley et al., 1980) mede o quanto o coeficiente de regressão se altera, em unidades de desvio padrão, quando a  $l$ -ésima observação é removida. Esta medida é calculada da seguinte forma:

$$\text{Dfbetas}_{m,l} = \frac{\hat{\beta}_m - \hat{\beta}_{m(l)}}{\sqrt{S_{(l)}^2 c_{mm}}},$$

em que  $\hat{\beta}_{m(l)}$  é o  $m$ -ésimo coeficiente estimado, quando a  $l$ -ésima observação é removida. Um valor será considerado influente se  $\text{Dfbetas}_{m,l} > 2/\sqrt{n}$ .



# Pontos influentes

O **Dffits** (*difference in fit(s)*, Belsley et al., 1980) mede a influência da  $l$ -ésima observação no valor ajustado, em unidades de desvio padrão. Esta medida é calculada da seguinte forma:

$$\text{Dffits}_l = \frac{\hat{Y}_l - \hat{Y}_{(l)}}{\sqrt{S_{(l)}^2 h_{ll}}}$$



# Pontos influentes

O **Dffits** (*difference in fit(s)*, Belsley et al., 1980) mede a influência da  $\ell$ -ésima observação no valor ajustado, em unidades de desvio padrão. Esta medida é calculada da seguinte forma:

$$\text{Dffits}_\ell = \frac{\hat{Y}_\ell - \hat{Y}_{(\ell)}}{\sqrt{S_{(\ell)}^2 h_{\ell\ell}}}.$$

Um valor será considerado influente se  $\text{Dffits}_\ell > 2\sqrt{p/n}$ .



# Pontos influentes

O **Dffits** (*difference in fit(s)*, Belsley et al., 1980) mede a influência da  $\ell$ -ésima observação no valor ajustado, em unidades de desvio padrão. Esta medida é calculada da seguinte forma:

$$\text{Dffits}_\ell = \frac{\hat{Y}_\ell - \hat{Y}_{(\ell)}}{\sqrt{S_{(\ell)}^2 h_{\ell\ell}}}.$$

Um valor será considerado influente se  $\text{Dffits}_\ell > 2\sqrt{p/n}$ . Dfbetas e Dffits são medidas equivalentes, ver Montgomery et al. (2021).



# Pontos influentes

O **Dffits** (*difference in fit(s)*, Belsley et al., 1980) mede a influência da  $\ell$ -ésima observação no valor ajustado, em unidades de desvio padrão. Esta medida é calculada da seguinte forma:

$$\text{Dffits}_\ell = \frac{\hat{Y}_\ell - \hat{Y}_{(\ell)}}{\sqrt{S_{(\ell)}^2 h_{\ell\ell}}}.$$

Um valor será considerado influente se  $\text{Dffits}_\ell > 2\sqrt{p/n}$ . Dfbetas e Dffits são medidas equivalentes, ver Montgomery et al. (2021).



# Pontos influentes

A **Covratio** traz informações sobre a precisão geral da estimativa. Esta medida é calculada da seguinte forma:

$$\text{Covratio}_\ell = \frac{|(\mathbf{X}_\ell^\top \mathbf{X}_\ell)^{-1} S_{(\ell)}^2|}{|(\mathbf{X}^\top \mathbf{X})^{-1} \text{QMRes}|} = \left[ \frac{S_{(\ell)}^2}{\text{QMRes}} \right]^p \frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$



# Pontos influentes

A **Covratio** traz informações sobre a precisão geral da estimativa. Esta medida é calculada da seguinte forma:

$$\text{Covratio}_\ell = \frac{|(\mathbf{X}_\ell^\top \mathbf{X}_\ell)^{-1} S_{(\ell)}^2|}{|(\mathbf{X}^\top \mathbf{X})^{-1} \text{QMRes}|} = \left[ \frac{S_{(\ell)}^2}{\text{QMRes}} \right]^p \frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

Se  $\text{Covratio}_\ell > 1$ , a  $\ell$ -ésima observação melhora a precisão do modelo mas, se  $\text{Covratio}_\ell < 1$ , ocorrerá o contrário.



# Pontos influentes

A **Covratio** traz informações sobre a precisão geral da estimativa. Esta medida é calculada da seguinte forma:

$$\text{Covratio}_\ell = \frac{|(\mathbf{X}_\ell^\top \mathbf{X}_\ell)^{-1} S_{(\ell)}^2|}{|(\mathbf{X}^\top \mathbf{X})^{-1} \text{QMRes}|} = \left[ \frac{S_{(\ell)}^2}{\text{QMRes}} \right]^p \frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

Se  $\text{Covratio}_\ell > 1$ , a  $\ell$ -ésima observação melhora a precisão do modelo mas, se  $\text{Covratio}_\ell < 1$ , ocorrerá o contrário. Uma alternativa é utilizar  $1 + 3p/n$  e  $1 - 3p/n$  como pontos de corte.



# Pontos influentes

A **Covratio** traz informações sobre a precisão geral da estimativa. Esta medida é calculada da seguinte forma:

$$\text{Covratio}_\ell = \frac{|(\mathbf{X}_\ell^\top \mathbf{X}_\ell)^{-1} S_{(\ell)}^2|}{|(\mathbf{X}^\top \mathbf{X})^{-1} \text{QMRes}|} = \left[ \frac{S_{(\ell)}^2}{\text{QMRes}} \right]^p \frac{h_{\ell\ell}}{1 - h_{\ell\ell}}.$$

Se  $\text{Covratio}_\ell > 1$ , a  $\ell$ -ésima observação melhora a precisão do modelo mas, se  $\text{Covratio}_\ell < 1$ , ocorrerá o contrário. Uma alternativa é utilizar  $1 + 3p/n$  e  $1 - 3p/n$  como pontos de corte.



# Pontos influentes

Uma vez que foi identificado observações influentes, elas poderão ser removidas caso sejam um erro de mensuração ou não pertençam a população de interesse.

Caso contrário, elas deverão ser mantidas e a adoção de técnicas robustas é necessária, por exemplo, metodologias que ponderam as observações.



# Pontos influentes

Uma vez que foi identificado observações influentes, elas poderão ser removidas caso sejam um erro de mensuração ou não pertençam a população de interesse.

Caso contrário, elas deverão ser mantidas e a adoção de técnicas robustas é necessária, por exemplo, metodologias que ponderam as observações.



# Roteiro

- 1 Introdução
- 2 Pontos de alavanca
- 3 Pontos influentes
- 4 Aplicação**
- 5 Referências bibliográficas



# 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,

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

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



# 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,

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

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



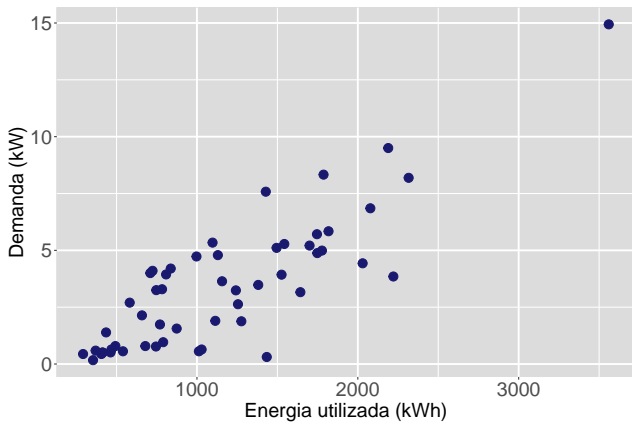


Figura 1: Gráfico de dispersão.

# Exemplo

Nós temos também que:

Tabela 1: Estimativas do parâmetros.

Parâmetro	Estimativa	EP	$t_c$
$\beta_1$	-0,831	0,442	-1,882
$\beta_2$	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 1: Estimativas do parâmetros.

Parâmetro	Estimativa	EP	$t_c$
$\beta_1$	-0,831	0,442	-1,882
$\beta_2$	0,004	0,001	11,030

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



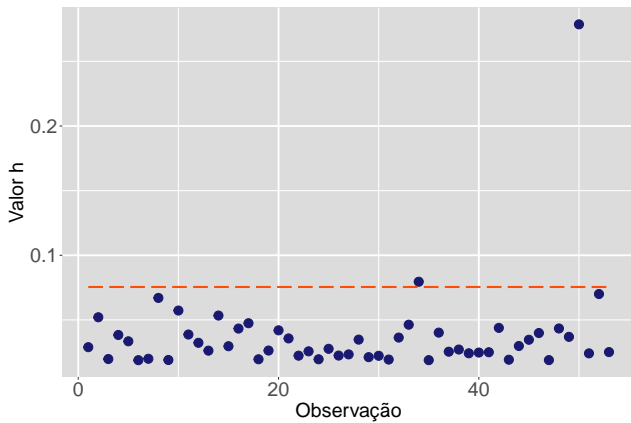


Figura 2: Gráfico de alavanca.

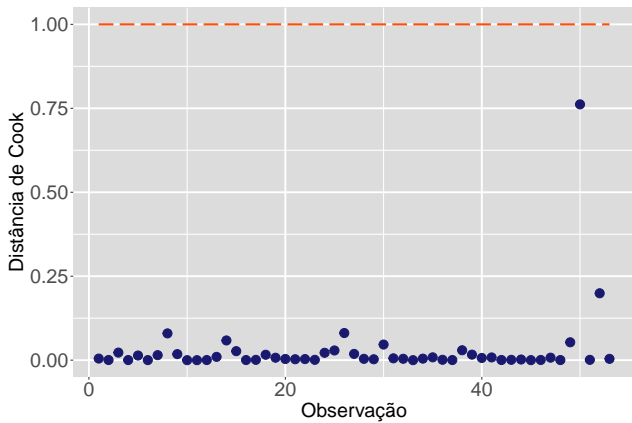
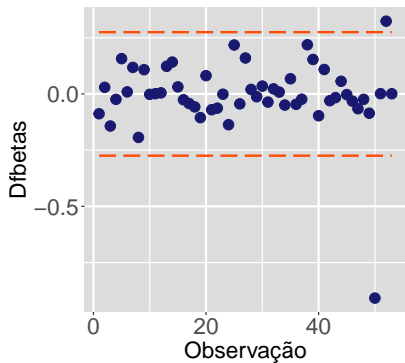
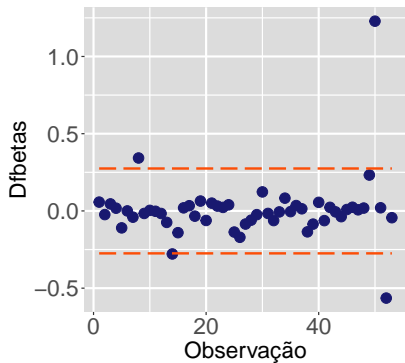


Figura 3: Distância de Cook.



(a) Intercepto.



(b) Demanda.

Figura 4: Dfbetas.

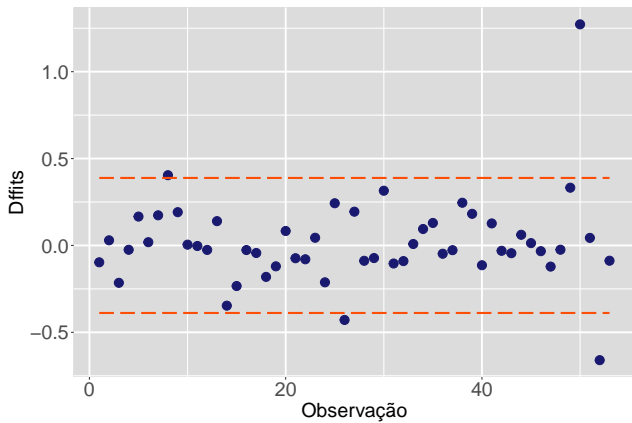


Figura 5: Dffits.

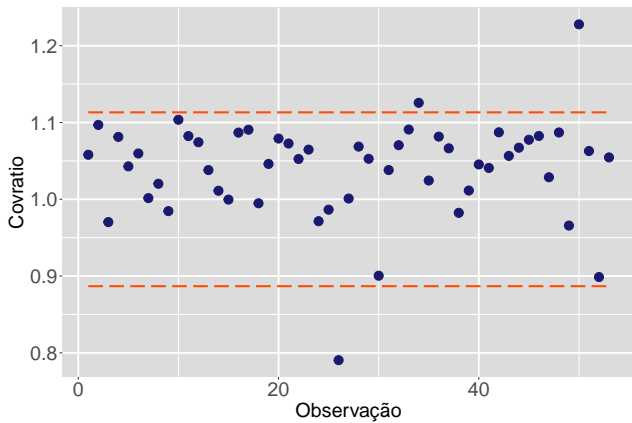


Figura 6: Covratio.

# Roteiro

- 1 Introdução
- 2 Pontos de alavanca
- 3 Pontos influentes
- 4 Aplicação
- 5 Referências bibliográficas**



# Referências bibliográficas I

- Belsley, D. A., Kuh, E. e Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.
- Cook, R. D. (1977), 'Detection of influential observation in linear regression', *Technometrics* **19**(1), 15–18.
- Cook, R. D. (1979), 'Influential observations in linear regression', *Journal of the American Statistical Association* **74**(365), 169–174.
- 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](mailto:tiago.magalhaes@ufjf.br)

📄 [ufjf.br/tiago\\_magalhaes](http://ufjf.br/tiago_magalhaes)

🌐 Dept. de Estatística, Sala 319

