Análise de resíduos

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 27 de abril de 2026

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, $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 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, $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2, também desconhecida, a ser estimada.

Forma matricial

A Equação (1) é o que nós definimos como **modelo de regressão normal linear** (MNL), podendo ser escrita de forma matricial, da seguinte forma:

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

Forma matricial

A Equação (1) é o que nós definimos como **modelo de regressão normal linear** (MNL), podendo ser escrita de forma matricial, da seguinte forma:

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

em que $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ e $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, sendo $\mathbf{0}$ o vetor nulo de dimensão n, $\boldsymbol{I}_n$ a matriz identidade de ordem n e $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$, a matriz de planejamento.

Forma matricial

A Equação (1) é o que nós definimos como **modelo de regressão normal linear** (MNL), podendo ser escrita de forma matricial, da seguinte forma:

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

em que $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ e $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, sendo $\mathbf{0}$ o vetor nulo de dimensão n, $\boldsymbol{I}_n$ a matriz identidade de ordem n e $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$, a matriz de planejamento.

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;
- O erro tem média zero;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- O erro tem média zero;
- O erro tem variância constante;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- O erro tem média zero;
- O erro tem variância constante;
- Os erros não são correlacionados;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- O erro tem média zero;
- O erro tem variância constante;
- Os erros não são correlacionados;
- Os erros têm distribuição normal, para procedimentos inferenciais.

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- O erro tem média zero;
- O erro tem variância constante;
- Os erros não são correlacionados;
- Os erros têm distribuição normal, para procedimentos inferenciais.

Método de máxima verossimilhança

O **estimador de máxima verossimilhança** (EMV) de $\boldsymbol{\beta}$ e σ^2 são dados, respectivamente, por:

$$\widehat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \tag{3}$$

$$\hat{\sigma}^2_{\mathrm{MV}} = \frac{1}{n}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}\right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}\right).$$

Método de máxima verossimilhança

O **estimador de máxima verossimilhança** (EMV) de β e σ^2 são dados, respectivamente, por:

$$\widehat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \tag{3}$$

$$\hat{\sigma}^2_{\mathrm{MV}} = \frac{1}{n} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta} \right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta} \right).$$

Observação: o EMV de σ^2 também pode ser escrito como função do quadrado médio do resíduo (QMRes).

Método de máxima verossimilhança

O **estimador de máxima verossimilhança** (EMV) de β e σ^2 são dados, respectivamente, por:

$$\widehat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \tag{3}$$

$$\hat{\sigma}^2_{\mathrm{MV}} = \frac{1}{n} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta} \right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta} \right).$$

Observação: o EMV de σ^2 também pode ser escrito como função do quadrado médio do resíduo (QMRes).

Método de máxima verossimilhança

Sob as condições de regularidades, nós temos que:

$$\widehat{\beta} \sim \mathcal{N}_p \left\{ \beta, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \right\}, \tag{4}$$

$$\hat{\sigma}^2_{\mathrm{MV}} \sim \mathcal{N} \left\{ \sigma^2, \frac{2(\sigma^2)^2}{n} \right\},$$

Método de máxima verossimilhança

Sob as condições de regularidades, nós temos que:

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{p}\left\{\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right\}, \tag{4}$$

$$\hat{\sigma}^2_{\mathrm{MV}} \sim \mathcal{N}\left\{\sigma^2, \frac{2(\sigma^2)^2}{n}\right\},$$

quando o tamanho de amostra é grande, adicionalmente, $\widehat{\boldsymbol{\beta}}$ e $\hat{\sigma}^2_{\mathrm{MV}}$ são ortogonais.

Método de máxima verossimilhança

Sob as condições de regularidades, nós temos que:

$$\widehat{\beta} \sim \mathcal{N}_p \left\{ \beta, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \right\}, \tag{4}$$

$$\hat{\sigma}^2_{\mathrm{MV}} \sim \mathcal{N} \left\{ \sigma^2, \frac{2(\sigma^2)^2}{n} \right\},$$

quando o tamanho de amostra é grande, adicionalmente, $\widehat{\beta}$ e $\hat{\sigma}^2_{\mathrm{MV}}$ são ortogonais.

Roteiro

Resíduos

Nós definimos como o ℓ-ésimo resíduo a diferença,

$$e_\ell = Y_\ell - \hat{Y}_\ell,$$

$\ell = 1, 2 \ldots, n$. Com nós vimos anteriormente, o resíduo pode ser expresso de forma matricial:

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y},$$

em que $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^\top$.

Resíduos

Nós definimos como o ℓ-ésimo resíduo a diferença,

$$e_\ell = Y_\ell - \hat{Y}_\ell,$$

$\ell = 1, 2 \ldots, n$. Com nós vimos anteriormente, o resíduo pode ser expresso de forma matricial:

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y},$$

em que $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^\top$.

Resíduos

A esperança (o vetor de esperanças) e a variância (matriz de covariâncias) de $\boldsymbol{e}$ são dada, respectivamente, por

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

Resíduos

A esperança (o vetor de esperanças) e a variância (matriz de covariâncias) de $\boldsymbol{e}$ são dada, respectivamente, por

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

Seja h_{ij}, o (i, j)-ésimo termo da matriz $\boldsymbol{H}$, então nós temos que: $\mathbb{E}(e_i) = 0$, $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$ e $\mathrm{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$, $i, j = 1, 2, \ldots, n$.

Resíduos

A esperança (o vetor de esperanças) e a variância (matriz de covariâncias) de $\boldsymbol{e}$ são dada, respectivamente, por

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

Seja h_{ij}, o (i,j)-ésimo termo da matriz $\boldsymbol{H}$, então nós temos que: $\mathbb{E}(e_i) = 0$, $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$ e $\mathrm{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$, $i, j = 1, 2, \ldots, n$.

Resíduos

Adicionalmente, nós podemos definir uma aproximação para a variância média, da seguinte forma:

$$\sum_{\ell=1}^{n} \frac{(e_\ell - \bar{e})^2}{n-p} = \sum_{\ell=1}^{n} \frac{e_\ell^2}{n-p} = \text{QMRes}.$$

Resíduos

Adicionalmente, nós podemos definir uma aproximação para a variância média, da seguinte forma:

$$\sum_{\ell=1}^{n} \frac{(e_\ell - \bar{e})^2}{n-p} = \sum_{\ell=1}^{n} \frac{e_\ell^2}{n-p} = \text{QMRes}.$$

Roteiro

Padronização de resíduos

Padronizar os resíduos nos ajuda a identificar valores atípicos (*outliers*) e valores extremos. Nós temos quatro tipos de resíduos padronizados:

1. Resíduos semi-estudentizados;

Padronização de resíduos

Padronizar os resíduos nos ajuda a identificar valores atípicos (*outliers*) e valores extremos. Nós temos quatro tipos de resíduos padronizados:

1. Resíduos semi-estudentizados;
2. Resíduos estudentizados;

Padronização de resíduos

Padronizar os resíduos nos ajuda a identificar valores atípicos (*outliers*) e valores extremos. Nós temos quatro tipos de resíduos padronizados:

1. Resíduos semi-estudentizados;
2. Resíduos estudentizados;
3. Resíduos *PRESS*;

Padronização de resíduos

Padronizar os resíduos nos ajuda a identificar valores atípicos (*outliers*) e valores extremos. Nós temos quatro tipos de resíduos padronizados:

1. Resíduos semi-estudentizados;
2. Resíduos estudentizados;
3. Resíduos *PRESS*;
4. Resíduos R-student.

Padronização de resíduos

Padronizar os resíduos nos ajuda a identificar valores atípicos (*outliers*) e valores extremos. Nós temos quatro tipos de resíduos padronizados:

1. Resíduos semi-estudentizados;
2. Resíduos estudentizados;
3. Resíduos *PRESS*;
4. Resíduos R-student.

Padronização de resíduos

Os **resíduos semi-estudentizados** ou, simplesmente, denominados de resíduos padronizados são expressos da seguinte forma:

$$d_\ell = \frac{e_\ell}{\sqrt{\sigma^2}}, \; \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

Os **resíduos semi-estudentizados** ou, simplesmente, denominados de resíduos padronizados são expressos da seguinte forma:

$$d_\ell = \frac{e_\ell}{\sqrt{\sigma^2}}, \ \ell = 1, 2, \ldots, n.$$

O parâmetro σ^2 pode ser estimado pelo QMRes. d_ℓ tem média zero e variância, aproximadamente, igual a 1. Valores “grandes” de d_ℓ ($|d_\ell| > 3$) podem indicar a presença de valores atípicos.

Padronização de resíduos

Os **resíduos semi-estudentizados** ou, simplesmente, denominados de resíduos padronizados são expressos da seguinte forma:

$$d_\ell = \frac{e_\ell}{\sqrt{\sigma^2}},\ \ell = 1, 2, \ldots, n.$$

O parâmetro σ^2 pode ser estimado pelo QMRes. d_ℓ tem média zero e variância, aproximadamente, igual a 1. Valores "grandes" de d_ℓ ($|d_\ell| > 3$) podem indicar a presença de valores atípicos.

Padronização de resíduos

Os **resíduos estudentizados** são expressos da seguinte forma:

$$r_\ell = \frac{e_\ell}{\sqrt{\sigma^2(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

Os **resíduos estudentizados** são expressos da seguinte forma:

$$r_\ell = \frac{e_\ell}{\sqrt{\sigma^2(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

O parâmetro σ^2 pode ser estimado pelo QMRes. r_ℓ tem média zero e variância igual a 1. Geralmente, os valores de r_ℓ são maiores do que seus correspondentes d_ℓ.

Padronização de resíduos

Os **resíduos estudentizados** são expressos da seguinte forma:

$$r_\ell = \frac{e_\ell}{\sqrt{\sigma^2(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

O parâmetro σ^2 pode ser estimado pelo QMRes. r_ℓ tem média zero e variância igual a 1. Geralmente, os valores de r_ℓ são maiores do que seus correspondentes d_ℓ.

Padronização de resíduos

Seja o ℓ-ésimo **resíduo PRESS** (*prediction error sum of squares*)

$$e_{(\ell)} = Y_\ell - \hat{Y}_{(\ell)}, \ \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

Seja o ℓ-ésimo **resíduo PRESS** (*prediction error sum of squares*)

$$e_{(\ell)} = Y_\ell - \hat{Y}_{(\ell)},\ \ell = 1, 2, \ldots, n.$$

em que $\hat{Y}_{(\ell)}$ é a variável resposta ajustada sem a observação ℓ e a ℓ-ésima linha da matriz de planejamento. Esta definição de resíduos pode verificar o quanto incomum é a observação ℓ.

Padronização de resíduos

Seja o ℓ-ésimo **resíduo PRESS** (*prediction error sum of squares*)

$$e_{(\ell)} = Y_\ell - \hat{Y}_{(\ell)},\ \ell = 1, 2, \ldots, n.$$

em que $\hat{Y}_{(\ell)}$ é a variável resposta ajustada sem a observação ℓ e a ℓ-ésima linha da matriz de planejamento. Esta definição de resíduos pode verificar o quanto incomum é a observação ℓ.

Padronização de resíduos

O ℓ-ésimo resíduo PRESS pode ser escrito da seguinte forma (ver Montgomery et al., 2021, p. 619):

$$e_{(\ell)} = \frac{e_\ell}{1 - h_{\ell\ell}}, \; \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

O ℓ-ésimo resíduo PRESS pode ser escrito da seguinte forma (ver Montgomery et al., 2021, p. 619):

$$e_{(\ell)} = \frac{e_\ell}{1 - h_{\ell\ell}}, \ \ell = 1, 2, \ldots, n.$$

E não é difícil mostrar que:

$$\mathrm{Var}(e_{(\ell)}) = \frac{\sigma^2}{1 - h_{\ell\ell}}, \ \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

O ℓ-ésimo resíduo PRESS pode ser escrito da seguinte forma (ver Montgomery et al., 2021, p. 619):

$$e_{(\ell)} = \frac{e_\ell}{1 - h_{\ell\ell}}, \ \ell = 1, 2, \ldots, n.$$

E não é difícil mostrar que:

$$\text{Var}(e_{(\ell)}) = \frac{\sigma^2}{1 - h_{\ell\ell}}, \ \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

Os **resíduos PRESS padronizados** são expressos da seguinte forma:

$$\frac{e_{(\ell)}}{\sqrt{\mathrm{Var}(e_{(\ell)})}} = \frac{e_\ell}{\sqrt{\sigma^2(1-h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

Os **resíduos PRESS padronizados** são expressos da seguinte forma:

$$\frac{e_{(\ell)}}{\sqrt{\mathrm{Var}(e_{(\ell)})}} = \frac{e_\ell}{\sqrt{\sigma^2(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

Notem, os resíduos PRESS padronizados coincidem com os estudentizados.

Padronização de resíduos

Os **resíduos PRESS padronizados** são expressos da seguinte forma:

$$\frac{e_{(\ell)}}{\sqrt{\mathrm{Var}(e_{(\ell)})}} = \frac{e_\ell}{\sqrt{\sigma^2(1 - h_{\ell\ell})}}, \;\; \ell = 1, 2, \ldots, n.$$

Notem, os resíduos PRESS padronizados coincidem com os estudentizados.

Padronização de resíduos

Como nós estamos levando em consideração o ajuste do modelo sem a ℓ-ésima observação, é natural nós pensarmos em um estimador para σ^2 que também desconsidere a observação ℓ. Esse estimador é dado por (ver Montgomery et al., 2021, p. 621):

$$S^2_{(\ell)} = \frac{(n-p)\text{QMRes} - \frac{e^2_\ell}{1-h_{\ell\ell}}}{n-p-1}, \ \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

Como nós estamos levando em consideração o ajuste do modelo sem a ℓ-ésima observação, é natural nós pensarmos em um estimador para σ^2 que também desconsidere a observação ℓ. Esse estimador é dado por (ver Montgomery et al., 2021, p. 621):

$$S^2_{(\ell)} = \frac{(n-p)\text{QMRes} - \frac{e_\ell^2}{1-h_{\ell\ell}}}{n-p-1},\ \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

O resíduo **R-student** é expresso da seguinte forma:

$$t_\ell = \frac{e_\ell}{\sqrt{S^2_{(\ell)}(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

Padronização de resíduos

O resíduo **R-student** é expresso da seguinte forma:

$$t_\ell = \frac{e_\ell}{\sqrt{S^2_{(\ell)}(1 - h_{\ell\ell})}}, \; \ell = 1, 2, \ldots, n.$$

Este resíduo também é denominado de resíduo estudentizado externamente.

Padronização de resíduos

O resíduo **R-student** é expresso da seguinte forma:

$$t_\ell = \frac{e_\ell}{\sqrt{S^2_{(\ell)}(1 - h_{\ell\ell})}}, \ \ell = 1, 2, \ldots, n.$$

Este resíduo também é denominado de resíduo estudentizado externamente.

Roteiro

Gráfico de resíduos

A disposição dos resíduos em um gráfico nos ajuda a observar o comportamento global deles e a identificar padrões.

Gráfico de resíduos

- Gráfico de probabilidade normal dos resíduos;

Gráfico de resíduos

- Gráfico de probabilidade normal dos resíduos;
  - Verifica a suposição de normalidade.

Gráfico de resíduos

- Gráfico de probabilidade normal dos resíduos;
  - Verifica a suposição de normalidade.
- Resíduos contra os valores ajustado;

Gráfico de resíduos

- Gráfico de probabilidade normal dos resíduos;
    - Verifica a suposição de normalidade.
- Resíduos contra os valores ajustado;
    - Verifica variância não constante;

Gráfico de resíduos

- Gráfico de probabilidade normal dos resíduos;
  - Verifica a suposição de normalidade.
- Resíduos contra os valores ajustado;
  - Verifica variância não constante;
  - Verifica não linearidade;

Gráfico de resíduos

- Gráfico de probabilidade normal dos resíduos;
  - Verifica a suposição de normalidade.
- Resíduos contra os valores ajustado;
  - Verifica variância não constante;
  - Verifica não linearidade;
  - Verifica possíveis valores atípicos.

Gráfico de resíduos

- Gráfico de probabilidade normal dos resíduos;
    - Verifica a suposição de normalidade.
- Resíduos contra os valores ajustado;
    - Verifica variância não constante;
    - Verifica não linearidade;
    - Verifica possíveis valores atípicos.

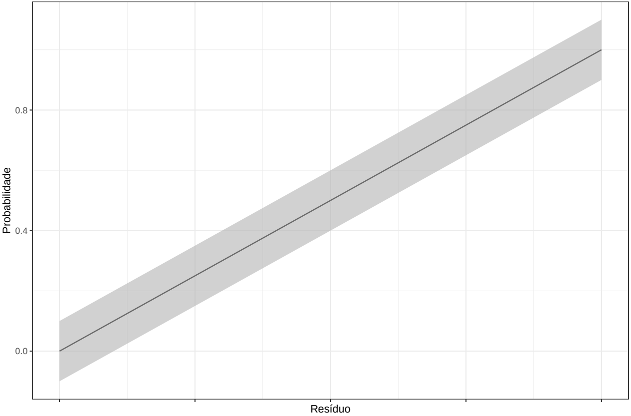
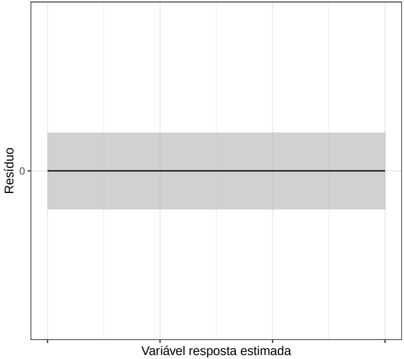
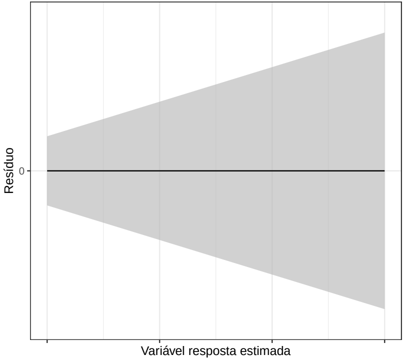


Figura 1: Gráfico de probabilidade normal dos resíduos.



(a) Variância constante.



(b) Variância não constante.

Figura 2: Gráficos de resíduos.

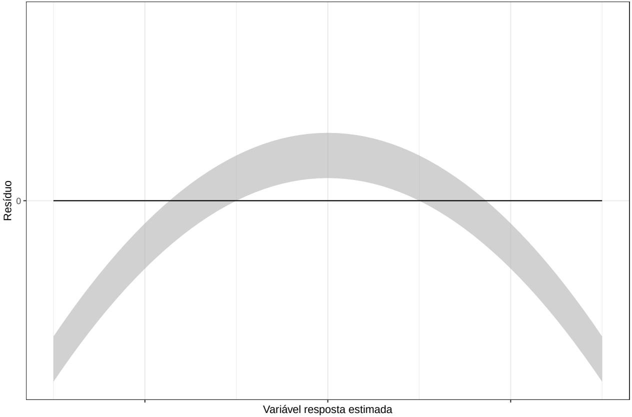


Figura 3: Não linearidade.

Gráfico de resíduos

- Resíduos contra as variáveis regressoras do modelo;

Gráfico de resíduos

- Resíduos contra as variáveis regressoras do modelo;
  - Verifica variância não constante;

Gráfico de resíduos

- Resíduos contra as variáveis regressoras do modelo;
    - Verifica variância não constante;
    - Verifica não linearidade.

Gráfico de resíduos

- Resíduos contra as variáveis regressoras do modelo;
  - Verifica variância não constante;
  - Verifica não linearidade.
- Resíduos contra as variáveis regressoras que não estão modelo;

Gráfico de resíduos

- Resíduos contra as variáveis regressoras do modelo;
  - Verifica variância não constante;
  - Verifica não linearidade.
- Resíduos contra as variáveis regressoras que não estão modelo;
  - Se um padrão aparecer, pode indicar que adicionar esse regressor pode melhorar o ajuste do modelo.

Gráfico de resíduos

- Resíduos contra as variáveis regressoras do modelo;
  - Verifica variância não constante;
  - Verifica não linearidade.
- Resíduos contra as variáveis regressoras que não estão modelo;
  - Se um padrão aparecer, pode indicar que adicionar esse regressor pode melhorar o ajuste do modelo.
- Resíduos contra a ordem das observações.

Gráfico de resíduos

- Resíduos contra as variáveis regressoras do modelo;
  - Verifica variância não constante;
  - Verifica não linearidade.
- Resíduos contra as variáveis regressoras que não estão modelo;
  - Se um padrão aparecer, pode indicar que adicionar esse regressor pode melhorar o ajuste do modelo.
- Resíduos contra a ordem das observações.
  - Verifica erros correlacionados.

Gráfico de resíduos

- Resíduos contra as variáveis regressoras do modelo;
  - Verifica variância não constante;
  - Verifica não linearidade.
- Resíduos contra as variáveis regressoras que não estão modelo;
  - Se um padrão aparecer, pode indicar que adicionar esse regressor pode melhorar o ajuste do modelo.
- Resíduos contra a ordem das observações.
  - Verifica erros correlacionados.

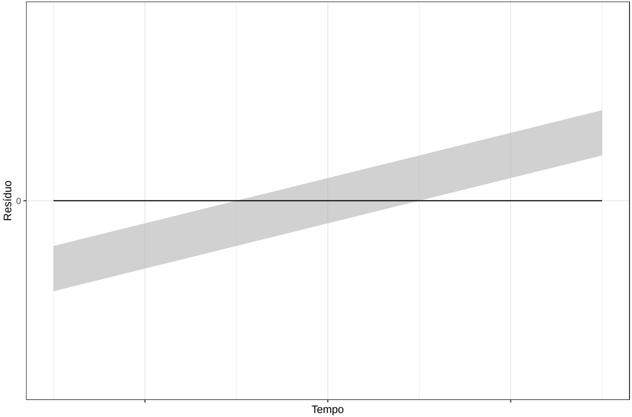


Figura 4: Não independência dos erros.

Gráficos de regressão parcial

Uma alternativa para nós verificarmos a contribuição marginal da variável independente x_m, $m = 1, 2, \ldots, p$, no modelo ajustado com as demais variáveis, é através dos gráficos de regressão parcial. Além disso, esse método também pode ser utilizado para verificar a relação correta entre Y e x_m.

Gráficos de regressão parcial

O método consiste em ajustar um modelo de regressão sem a covariável x_m e calcular os resíduos. Em seguida, ajustar um modelo de regressão com x_m como variável resposta e a demais covariáveis como regressoras e calcular os resíduos dessa análise. E, finalmente, construir um gráfico de dispersão com estes dois conjuntos de resíduos.

Gráficos de regressão parcial

Se

- o gráfico parece ser linear, então uma relação linear entre Y e x_m parece ser razoável;

Gráficos de regressão parcial

Se

- o gráfico parece ser linear, então uma relação linear entre Y e x_m parece ser razoável;
- o gráfico for curvilíneo, pode ser necessário trabalhar com x_m^2 ou $1/x_m$;

Gráficos de regressão parcial

Se

- o gráfico parece ser linear, então uma relação linear entre Y e x_m parece ser razoável;
- o gráfico for curvilíneo, pode ser necessário trabalhar com x_m^2 ou $1/x_m$;
- x_m for uma variável candidata e uma "banda" horizontal aparecer, essa variável não adiciona nenhuma informação nova.

Gráficos de regressão parcial

Se

- o gráfico parece ser linear, então uma relação linear entre Y e x_m parece ser razoável;
- o gráfico for curvilíneo, pode ser necessário trabalhar com x_m^2 ou $1/x_m$;
- x_m for uma variável candidata e uma "banda" horizontal aparecer, essa variável não adiciona nenhuma informação nova.

Gráficos de regressão parcial

Considerações:

- Use com cautela, eles apenas sugerem possíveis relações;

Gráficos de regressão parcial

Considerações:

- Use com cautela, eles apenas sugerem possíveis relações;
- Geralmente não detectam efeitos de interação;

Gráficos de regressão parcial

Considerações:

- Use com cautela, eles apenas sugerem possíveis relações;
- Geralmente não detectam efeitos de interação;
- Se houver multicolinearidade, os gráficos de regressão podem fornecer informações incorretas.

Gráficos de regressão parcial

Considerações:

- Use com cautela, eles apenas sugerem possíveis relações;
- Geralmente não detectam efeitos de interação;
- Se houver multicolinearidade, os gráficos de regressão podem fornecer informações incorretas.

Outros gráficos

Construir um gráfico de dispersão para todos os pares de covariáveis, isto pode dar informações sobre a relação entre elas, como:

- uma possível correlação;

Outros gráficos

Construir um gráfico de dispersão para todos os pares de covariáveis, isto pode dar informações sobre a relação entre elas, como:

- uma possível correlação;
- a existência de "pontos remotos".

Outros gráficos

Construir um gráfico de dispersão para todos os pares de covariáveis, isto pode dar informações sobre a relação entre elas, como:

- uma possível correlação;
- a existência de "pontos remotos".

Roteiro

Estatística PRESS

Uma valiosa maneira para comparação de modelos é a estatística PRESS (Montgomery et al., 2021), ela é definida da seguinte forma

$$\text{PRESS} = \sum_{\ell=1}^{n} e_{(\ell)}^{2} = \sum_{\ell=1}^{n} \left\{ Y_{\ell} - \hat{Y}_{(\ell)} \right\}^{2} = \sum_{\ell=1}^{n} \left\{ \frac{e_{\ell}}{1 - h_{\ell\ell}} \right\}^{2}.$$

Estatística PRESS

Uma valiosa maneira para comparação de modelos é a estatística PRESS (Montgomery et al., 2021), ela é definida da seguinte forma

$$\text{PRESS} = \sum_{\ell=1}^{n} e_{(\ell)}^2 = \sum_{\ell=1}^{n} \left\{ Y_\ell - \hat{Y}_{(\ell)} \right\}^2 = \sum_{\ell=1}^{n} \left\{ \frac{e_\ell}{1 - h_{\ell\ell}} \right\}^2 .$$

Observação: valores pequenos da estatística PRESS são desejados.

Estatística PRESS

Uma valiosa maneira para comparação de modelos é a estatística PRESS (Montgomery et al., 2021), ela é definida da seguinte forma

$$\text{PRESS} = \sum_{\ell=1}^{n} e_{(\ell)}^2 = \sum_{\ell=1}^{n} \left\{ Y_\ell - \hat{Y}_{(\ell)} \right\}^2 = \sum_{\ell=1}^{n} \left\{ \frac{e_\ell}{1 - h_{\ell\ell}} \right\}^2.$$

Observação: valores pequenos da estatística PRESS são desejados.

Estatística PRESS

Também é possível definir um coeficiente de determinação da estatística PRESS da seguinte forma,

$$R^2 = 1 - \frac{\text{PRESS}}{\text{SQT}}.$$

Estatística PRESS

Também é possível definir um coeficiente de determinação da estatística PRESS da seguinte forma,

$$R^2 = 1 - \frac{\text{PRESS}}{\text{SQT}}.$$

Interpretação: nós esperamos que o modelo explique cerca de R^2% da variabilidade na previsão de uma nova observação.

Estatística PRESS

Também é possível definir um coeficiente de determinação da estatística PRESS da seguinte forma,

$$R^2 = 1 - \frac{\text{PRESS}}{\text{SQT}}.$$

Interpretação: nós esperamos que o modelo explique cerca de $R^2\%$ da variabilidade na previsão de uma nova observação.

Roteiro

Aplicação

(Montgomery et al., 2021, p. 76) Um conjunto de dados que relaciona o tempo de entrega de máquinas de venda automática (Y, em minutos) com o número de máquinas em estoque (x_2) e o comprimento da rota (x_3, em pés). Após o ajuste, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = 2{,}341 + 1{,}661 x_{\ell 2} + 0{,}014 x_{\ell 3},$$

$\ell = 1, 2, \ldots, 25.$

Aplicação

(Montgomery et al., 2021, p. 76) Um conjunto de dados que relaciona o tempo de entrega de máquinas de venda automática (Y, em minutos) com o número de máquinas em estoque (x_2) e o comprimento da rota (x_3, em pés). Após o ajuste, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = 2{,}341 + 1{,}661 x_{\ell 2} + 0{,}014 x_{\ell 3},$$

$\ell = 1, 2, \ldots, 25$.

Exemplo

Nós temos também que:

Tabela 1: Estimativas do parâmetros.

Parâmetro	Estimativa	EP	t_c
β_1	2,341	1,097	2,135
β_2	1,616	0,171	9,464
β_3	0,014	0,004	3,981

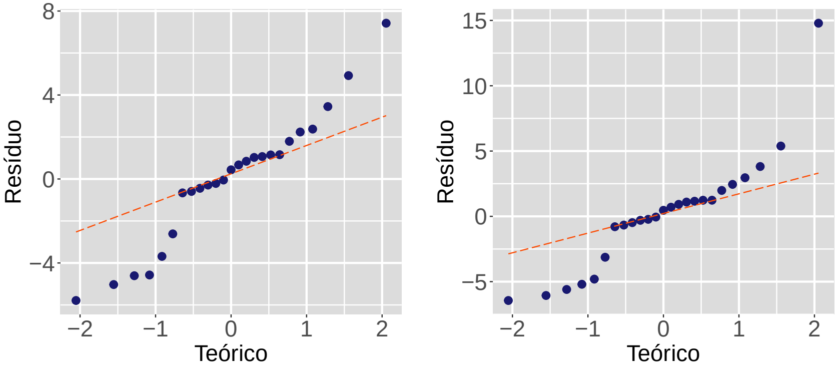
Região crítica, para $\alpha = 5\%$: $|t_c| > 2{,}074$ com QMRes $= 10{,}164$.

Nós temos também que:

Tabela 1: Estimativas do parâmetros.

Parâmetro	Estimativa	EP	t_c
β_1	2,341	1,097	2,135
β_2	1,616	0,171	9,464
β_3	0,014	0,004	3,981

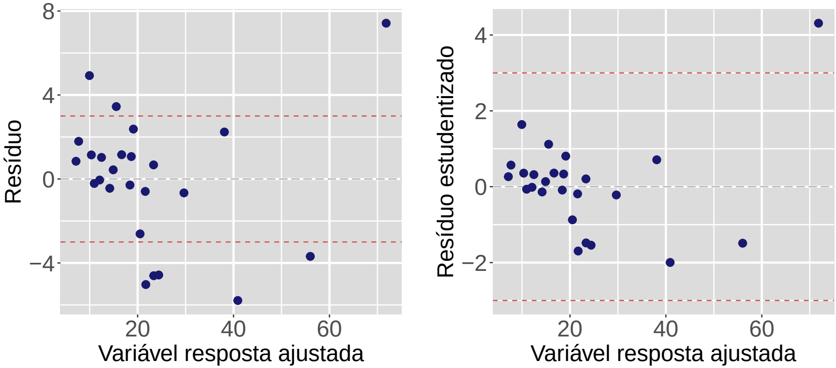
Região crítica, para $\alpha = 5\%$: $|t_c| > 2{,}074$ com QMRes $= 10{,}164$.



(a) Resíduos ordinários.

(b) Resíduos PRESS.

Figura 5: Gráficos QQ.



(a) Resíduos ordinários. (b) Resíduos estudentizados.

Figura 6: Gráficos de resíduos.

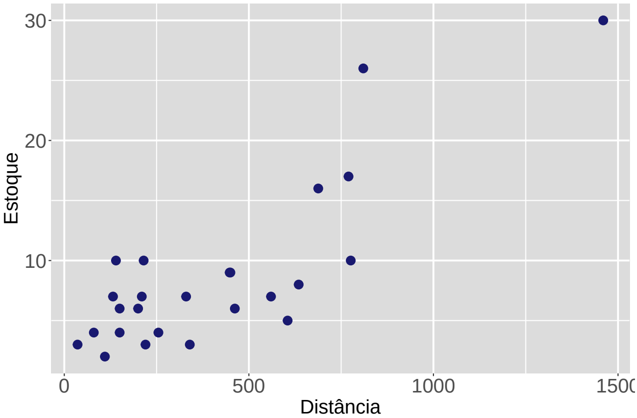


Figura 7: Gráfico de dispersão entre as covariáveis

Roteiro

Referências bibliográficas I

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