

Validação

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 08 de julho de 2026



Roteiro

- 1 Introdução
- 2 Técnicas de validação
- 3 Aplicações
- 4 Referências bibliográficas



Roteiro

- 1 Introdução
- 2 Técnicas de validação
- 3 Aplicações
- 4 Referências bibliográficas



Modelo de regressão linear

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

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



Modelo de regressão linear

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

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

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 σ^2 , também desconhecida, a ser estimada.



Modelo de regressão linear

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

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

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 σ^2 , também desconhecida, a ser estimada.



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.



Validação de modelos

Após a seleção de um modelo final, a última parte do processo de construção de modelos é a **validação**.

Esta é a etapa onde nós observamos se o modelo se comporta bem ou funciona conforme pretendido (no caso, para situações em que há um conhecimento prévio).



Validação de modelos

Após a seleção de um modelo final, a última parte do processo de construção de modelos é a **validação**.

Esta é a etapa onde nós observamos se o modelo se comporta bem ou funciona conforme pretendido (no caso, para situações em que há um conhecimento prévio).



Validação de modelos

Quando não há um conhecimento prévio, nós verificamos o comportamento dos coeficientes estimados e dos valores preditos pelo modelo nos seguintes pontos:

- se o sinais dos coeficientes são “inapropriados”;



Validação de modelos

Quando não há um conhecimento prévio, nós verificamos o comportamento dos coeficientes estimados e dos valores preditos pelo modelo nos seguintes pontos:

- se o sinais dos coeficientes são “inapropriados”;
- se a magnitude dos coeficientes estão em acordo com os dados;



Validação de modelos

Quando não há um conhecimento prévio, nós verificamos o comportamento dos coeficientes estimados e dos valores preditos pelo modelo nos seguintes pontos:

- se o sinais dos coeficientes são “inapropriados”;
- se a magnitude dos coeficientes estão em acordo com os dados;
- a estabilidade das estimativas dos coeficientes;



Validação de modelos

Quando não há um conhecimento prévio, nós verificamos o comportamento dos coeficientes estimados e dos valores preditos pelo modelo nos seguintes pontos:

- se o sinais dos coeficientes são “inapropriados”;
- se a magnitude dos coeficientes estão em acordo com os dados;
- a estabilidade das estimativas dos coeficientes;
- se os valores preditos estão de acordo com a natureza dos dados



Validação de modelos

Quando não há um conhecimento prévio, nós verificamos o comportamento dos coeficientes estimados e dos valores preditos pelo modelo nos seguintes pontos:

- se o sinais dos coeficientes são “inapropriados”;
- se a magnitude dos coeficientes estão em acordo com os dados;
- a estabilidade das estimativas dos coeficientes;
- se os valores preditos estão de acordo com a natureza dos dados.



Validação de modelos

De maneira geral, a melhor forma de validar um modelo é reestimá-lo a partir de um novo conjunto de dados.

Porém, devido as dificuldades associadas, se prefere dividir a amostra original em duas partes: **estimação** e **predição**.



Validação de modelos

De maneira geral, a melhor forma de validar um modelo é reestimá-lo a partir de um novo conjunto de dados.

Porém, devido as dificuldades associadas, se prefere dividir a amostra original em duas partes: **estimação** e **predição**.

O procedimento de particionar o banco de dados para poder ajustá-lo e validá-lo é chamado de **validação cruzada**.



Validação de modelos

De maneira geral, a melhor forma de validar um modelo é reestimá-lo a partir de um novo conjunto de dados.

Porém, devido as dificuldades associadas, se prefere dividir a amostra original em duas partes: **estimação** e **predição**.

O procedimento de particionar o banco de dados para poder ajustá-lo e validá-lo é chamado de **validação cruzada**.



Roteiro

- 1 Introdução
- 2 Técnicas de validação
- 3 Aplicações
- 4 Referências bibliográficas



Técnicas de validação

Seja a validação feita em uma base de dados obtida a partir da coleta de novas observações ou em uma amostra de predição, existem duas estratégias a serem adotadas:



Critérios para a seleção de modelos

1. A primeira consiste em ajustar na base de dados adicional um novo modelo contendo as mesmas variáveis selecionadas na base de dados original



Critérios para a seleção de modelos

1. A primeira consiste em ajustar na base de dados adicional um novo modelo contendo as mesmas variáveis selecionadas na base de dados original e comparar as estimativas dos parâmetros e medidas, como o R^2 e o AIC, do modelo original com o modelo na base nova;



Critérios para a seleção de modelos

1. A primeira consiste em ajustar na base de dados adicional um novo modelo contendo as mesmas variáveis selecionadas na base de dados original e comparar as estimativas dos parâmetros e medidas, como o R^2 e o AIC, do modelo original com o modelo na base nova;



Cr terios para a sele o de modelos

2. A segunda estrat gia consiste em prever o valor de Y para as observa es da base nova baseado nas estimativas dos par metros do modelo original



Cr terios para a sele o de modelos

2. A segunda estrat gia consiste em prever o valor de Y para as observa es da base nova baseado nas estimativas dos par metros do modelo original e comparar, por exemplo, a SQ_{Res} na base original com a SQ de predi o na base nova.



Critérios para a seleção de modelos

2. A segunda estratégia consiste em prever o valor de Y para as observações da base nova baseado nas estimativas dos parâmetros do modelo original e comparar, por exemplo, a SQ_{Res} na base original com a SQ de predição na base nova.

Roteiro

- 1 Introdução
- 2 Técnicas de validação
- 3 Aplicações**
- 4 Referências bibliográficas



Aplicação 1. (Hald, 1952) Um conjunto de dados, com 13 observações,

Aplicações

Aplicação 1. (Hald, 1952) Um conjunto de dados, com 13 observações, relacionando o calor transformado em calorias por grama de cimento (Y),



Aplicações

Aplicação 1. (Hald, 1952) Um conjunto de dados, com 13 observações, relacionando o calor transformado em calorias por grama de cimento (Y), com a quantidade de quatro tipos de mistura (x_2 a x_5).



Aplicações

Aplicação 1. (Hald, 1952) Um conjunto de dados, com 13 observações, relacionando o calor transformado em calorias por grama de cimento (Y), com a quantidade de quatro tipos de mistura (x_2 a x_5). Após uma análise de regressão, dois modelos foram propostos,



Aplicações

Aplicação 1. (Hald, 1952) Um conjunto de dados, com 13 observações, relacionando o calor transformado em calorias por grama de cimento (Y), com a quantidade de quatro tipos de mistura (x_2 a x_5). Após uma análise de regressão, dois modelos foram propostos,

$$M1 : \hat{Y}_\ell = 52,58 + 1,468x_{\ell 2} + 0,662x_{\ell 3},$$

$$M2 : \hat{Y}_\ell = 71,65 + 1,452x_{\ell 2} + 0,416x_{\ell 3} - 0,237x_{\ell 5},$$

$$\ell = 1, 2, \dots, 13.$$



Aplicações

Aplicação 1. (Hald, 1952) Um conjunto de dados, com 13 observações, relacionando o calor transformado em calorias por grama de cimento (Y), com a quantidade de quatro tipos de mistura (x_2 a x_5). Após uma análise de regressão, dois modelos foram propostos,

$$M1 : \hat{Y}_\ell = 52,58 + 1,468x_{\ell 2} + 0,662x_{\ell 3},$$

$$M2 : \hat{Y}_\ell = 71,65 + 1,452x_{\ell 2} + 0,416x_{\ell 3} - 0,237x_{\ell 5},$$

$$\ell = 1, 2, \dots, 13.$$



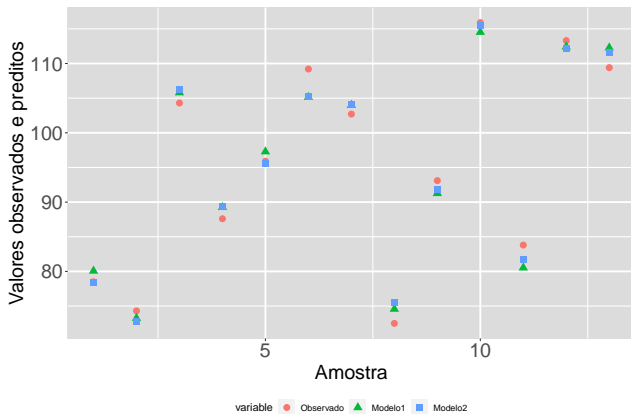


Figura 1: Valores observados e preditos pelos modelos.

Aplicação 2. (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)

Aplicação 2. (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).

Aplicação 2. (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,

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

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

Aplicação 2. (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,

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

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



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 $QMR_{es} = 10,164$. Agora, suponham que 15 novas observações foram coletadas.



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 $QMR_{es} = 10,164$. Agora, suponham que 15 novas observações foram coletadas.



Tabela 2: Amostra de predição.

Observado	Estimado	Diferença
51,00	50,91	0,09
16,80	21,13	-4,33
26,16	30,75	-4,59
19,90	17,61	2,29
24,00	26,42	-2,42
18,55	15,27	3,28
31,93	29,65	2,28
16,95	11,85	5,10
7,00	6,03	0,97
14,00	9,00	5,00
37,03	31,15	5,88
18,62	24,54	-5,92
16,10	15,81	0,29
24,38	20,45	3,93
64,75	76,06	-11,31

Exemplo

O erro de predição médio foi 0,035, o que pode ser considerado pequeno.

Nós temos também que, na amostra de predição,

$$\frac{\sum_{\ell=26}^{40} (y_{\ell} - \hat{y}_{\ell})^2}{15} = 22,122.$$

Exemplo

O erro de predição médio foi 0,035, o que pode ser considerado pequeno.

Nós temos também que, na amostra de predição,

$$\frac{\sum_{\ell=26}^{40} (y_{\ell} - \hat{y}_{\ell})^2}{15} = 22,122.$$

Como o $QMRes = 10,164$ é menor que o valor acima, o modelo não prevê novas observações tão bem como ele ajusta os dados existentes.



Exemplo

O erro de predição médio foi 0,035, o que pode ser considerado pequeno.

Nós temos também que, na amostra de predição,

$$\frac{\sum_{\ell=26}^{40} (y_{\ell} - \hat{y}_{\ell})^2}{15} = 22,122.$$

Como o $QMRes = 10,164$ é menor que o valor acima, o modelo não prevê novas observações tão bem como ele ajusta os dados existentes.



Exemplo

Calculando o coeficiente de determinação na amostra de predição, nós temos também que,

$$R_{\text{Pred}}^2 = 1 - \frac{\sum_{\ell=26}^{40} (y_{\ell} - \hat{y}_{\ell})^2}{\sum_{\ell=26}^{40} (y_{\ell} - \bar{y})^2} = 1 - \frac{331,83}{3206,23} = 0,8965.$$

Exemplo

Calculando o coeficiente de determinação na amostra de predição, nós temos também que,

$$R_{\text{Pred}}^2 = 1 - \frac{\sum_{\ell=26}^{40} (y_{\ell} - \hat{y}_{\ell})^2}{\sum_{\ell=26}^{40} (y_{\ell} - \bar{y})^2} = 1 - \frac{331,83}{3206,23} = 0,8965.$$

Como o $R^2 = 0,9596$ é maior que o valor acima, nós chegamos na mesma conclusão anterior, o modelo não prevê novas observações tão bem como ele ajusta os dados existentes.



Exemplo

Calculando o coeficiente de determinação na amostra de predição, nós temos também que,

$$R_{\text{Pred}}^2 = 1 - \frac{\sum_{\ell=26}^{40} (y_{\ell} - \hat{y}_{\ell})^2}{\sum_{\ell=26}^{40} (y_{\ell} - \bar{y})^2} = 1 - \frac{331,83}{3206,23} = 0,8965.$$

Como o $R^2 = 0,9596$ é maior que o valor acima, nós chegamos na mesma conclusão anterior, o modelo não prevê novas observações tão bem como ele ajusta os dados existentes.



Roteiro

- 1 Introdução
- 2 Técnicas de validação
- 3 Aplicações
- 4 Referências bibliográficas



Referências bibliográficas I

Hald, A. (1952), *Statistical theory with Engineering applications*, Wiley, New York.

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



The background features a stylized illustration. In the upper half, a city skyline with various skyscrapers is rendered in a light, sketchy style. Below the city, a dense forest of tall, thin evergreen trees is depicted. In the foreground, three bears—a brown bear, a black and white panda, and a white bear—are standing on a grassy hill, looking towards the right. A blue diamond-shaped kite with a red tail is flying in the sky, held by a string that extends from the white bear.

Obrigado!

✉ tiago.magalhaes@ufjf.br

🌐 ufjf.br/tiago_magalhaes

🌐 Departamento de Estatística, Sala 319