

Análise de variância

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 23 de abril de 2025



Roteiro

- 1 Introdução
- 2 Partição de somas de quadrados
- 3 Coeficiente de determinação
- 4 Aplicação
- 5 Estudo de simulação
- 6 Referências bibliográficas



Roteiro

- 1 Introdução
- 2 Partição de somas de quadrados
- 3 Coeficiente de determinação
- 4 Aplicação
- 5 Estudo de simulação
- 6 Referências bibliográficas



Modelo de regressão linear

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

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

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, $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2 , também desconhecida, a ser estimada. Nós vamos assumir que (1) é um modelo com intercepto, $x_{\ell 1} = 1$.



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, $\varepsilon_\ell \sim \mathcal{N}(0, \sigma^2)$, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2 , também desconhecida, a ser estimada. Nós vamos assumir que (1) é um modelo com intercepto, $x_{\ell 1} = 1$.



Forma matricial

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

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

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



Forma matricial

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

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

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



Introdução

Motivação

A análise de variância (*analysis of variance*, anova) é uma forma de nós vermos o modelo de regressão linear sob um outro tipo de ponto de vista.



Introdução

Motivação

A análise de variância (*analysis of variance*, anova) é uma forma de nós vermos o modelo de regressão linear sob um outro tipo de ponto de vista.



Roteiro

- 1 Introdução
- 2 Partição de somas de quadrados
- 3 Coeficiente de determinação
- 4 Aplicação
- 5 Estudo de simulação
- 6 Referências bibliográficas



Soma de quadrados total

O método da análise de variância é baseada na partição da somas de quadrados e dos graus de liberdade associados com a variável resposta Y .

Se todas as observações Y_ℓ são idênticas, não é necessário uma análise de regressão.



Soma de quadrados total

O método da análise de variância é baseada na partição da somas de quadrados e dos graus de liberdade associados com a variável resposta Y .

Se todas as observações Y_ℓ são idênticas, não é necessário uma análise de regressão. E a variação de Y_ℓ seria medida em termos dos desvios:

$$Y_\ell - \bar{Y}.$$

A Figura 1 apresenta esse desvio de forma gráfica.



Soma de quadrados total

O método da análise de variância é baseada na partição da somas de quadrados e dos graus de liberdade associados com a variável resposta Y .

Se todas as observações Y_ℓ são idênticas, não é necessário uma análise de regressão. E a variação de Y_ℓ seria medida em termos dos desvios:

$$Y_\ell - \bar{Y}.$$

A Figura 1 apresenta esse desvio de forma gráfica.



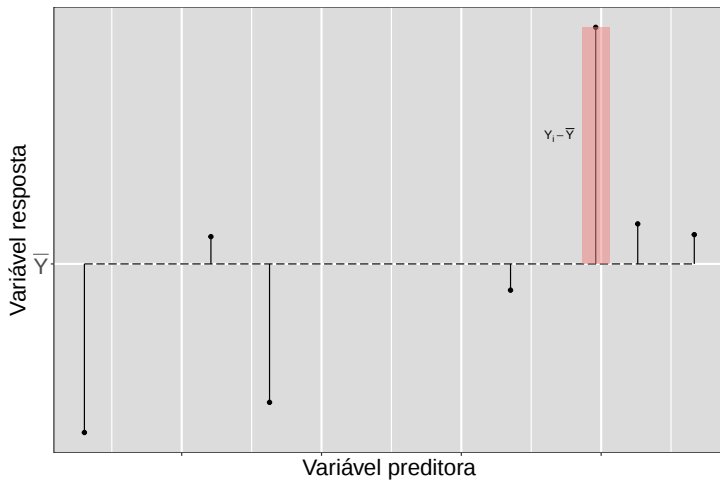


Figura 1: $Y_\ell - \bar{Y}$.

Soma de quadrados total

A medida da variação total dos desvios, a **soma de quadrados total** (SQT), é dada por

$$\text{SQT} = \sum_{\ell=1}^n \left(Y_{\ell} - \bar{Y} \right)^2.$$

Soma de quadrados total

A medida da variação total dos desvios, a **soma de quadrados total** (SQT), é dada por

$$\text{SQT} = \sum_{\ell=1}^n (Y_{\ell} - \bar{Y})^2.$$

Note que, se $\text{SQT} = 0$, todas as observações são iguais e quanto maior for SQT , maior é a variabilidade dos dados.

Soma de quadrados total

A medida da variação total dos desvios, a **soma de quadrados total** (SQT), é dada por

$$\text{SQT} = \sum_{\ell=1}^n (Y_{\ell} - \bar{Y})^2.$$

Note que, se $\text{SQT} = 0$, todas as observações são iguais e quanto maior for SQT , maior é a variabilidade dos dados.

Soma de quadrados dos resíduos

Como nós vimos na aula anterior, em uma análise de regressão, a variabilidade dos dados é medida em relação a reta de regressão, isto é,

$$Y_{\ell} - \hat{Y}_{\ell}.$$

A Figura 2 apresenta esse desvio de forma gráfica.



Soma de quadrados dos resíduos

Como nós vimos na aula anterior, em uma análise de regressão, a variabilidade dos dados é medida em relação a reta de regressão, isto é,

$$Y_{\ell} - \hat{Y}_{\ell}.$$

A Figura 2 apresenta esse desvio de forma gráfica.



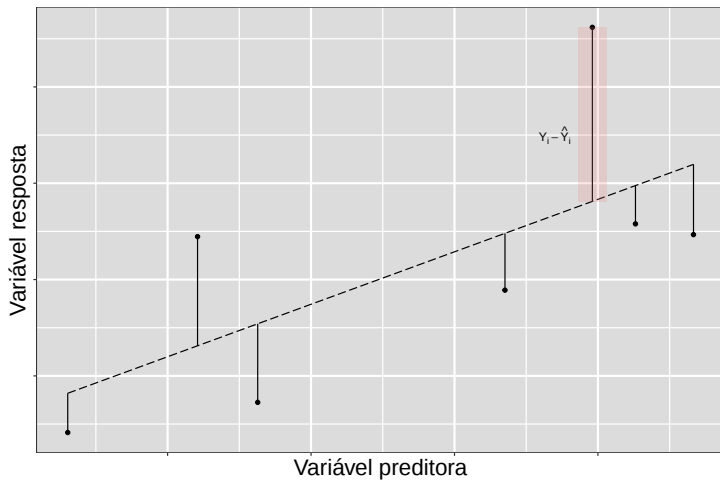


Figura 2: $Y_\ell - \hat{Y}_\ell$.

Soma de quadrados dos resíduos

A medida para avaliar toda a variação na amostra é a **soma de quadrado dos resíduos** (SQRes), dada por

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

Soma de quadrados dos resíduos

A medida para avaliar toda a variação na amostra é a **soma de quadrado dos resíduos** (SQRes), dada por

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

Se $\text{SQRes} = 0$, todas observações estão em cima da reta de regressão e quanto maior for SQRes, maior é a variabilidade dos dados em relação a reta de regressão.



Soma de quadrados dos resíduos

A medida para avaliar toda a variação na amostra é a **soma de quadrado dos resíduos** (SQRes), dada por

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

Se $\text{SQRes} = 0$, todas observações estão em cima da reta de regressão e quanto maior for SQRes, maior é a variabilidade dos dados em relação a reta de regressão.



Soma de quadrados da regressão

Também é possível comparar a distancia entre o valor estimado pela reta de regressão e a média, isto é,

$$\hat{Y}_\ell - \bar{Y}.$$

A Figura 3 apresenta esse desvio de forma gráfica.

Soma de quadrados da regressão

Também é possível comparar a distancia entre o valor estimado pela reta de regressão e a média, isto é,

$$\hat{Y}_\ell - \bar{Y}.$$

A Figura 3 apresenta esse desvio de forma gráfica.

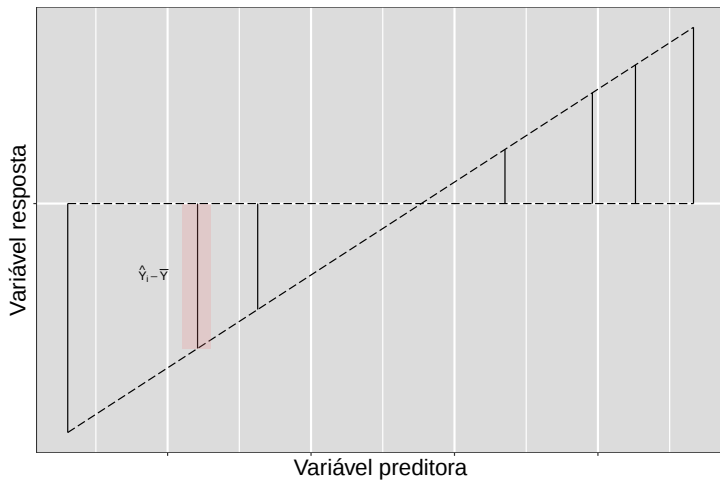


Figura 3: $\hat{Y}_\ell - \bar{Y}$.

Soma de quadrados da regressão

E, conseqüentemente, medir a distância quadrada entre elas, chamada de **soma de quadrados da regressão (SQReg)**,

$$\text{SQReg} = \sum_{\ell=1}^n \left(\hat{Y}_{\ell} - \bar{Y} \right)^2.$$

Soma de quadrados da regressão

E, conseqüentemente, medir a distância quadrada entre elas, chamada de **soma de quadrados da regressão** (SQReg),

$$\text{SQReg} = \sum_{\ell=1}^n \left(\hat{Y}_{\ell} - \bar{Y} \right)^2.$$

Se $\text{SQReg} = 0$, a reta de regressão é horizontal.

Soma de quadrados da regressão

E, conseqüentemente, medir a distância quadrada entre elas, chamada de **soma de quadrados da regressão** (SQReg),

$$\text{SQReg} = \sum_{\ell=1}^n \left(\hat{Y}_{\ell} - \bar{Y} \right)^2.$$

Se $\text{SQReg} = 0$, a reta de regressão é horizontal.

Soma de quadrados

Considere o desvio $Y_\ell - \bar{Y}$, a medida básica para mensurar a variação das observações Y_ℓ , a diferença entre o ℓ -ésimo valor da variável resposta e a média amostral pode ser decomposta como a soma de dois termos:

$$Y_\ell - \bar{Y} = (Y_\ell - \hat{Y}_\ell) + (\hat{Y}_\ell - \bar{Y}). \quad (3)$$

A Figura 4 apresenta esse desvio de forma gráfica.



Soma de quadrados

Considere o desvio $Y_\ell - \bar{Y}$, a medida básica para mensurar a variação das observações Y_ℓ , a diferença entre o ℓ -ésimo valor da variável resposta e a média amostral pode ser decomposta como a soma de dois termos:

$$Y_\ell - \bar{Y} = (Y_\ell - \hat{Y}_\ell) + (\hat{Y}_\ell - \bar{Y}). \quad (3)$$

A Figura 4 apresenta esse desvio de forma gráfica.



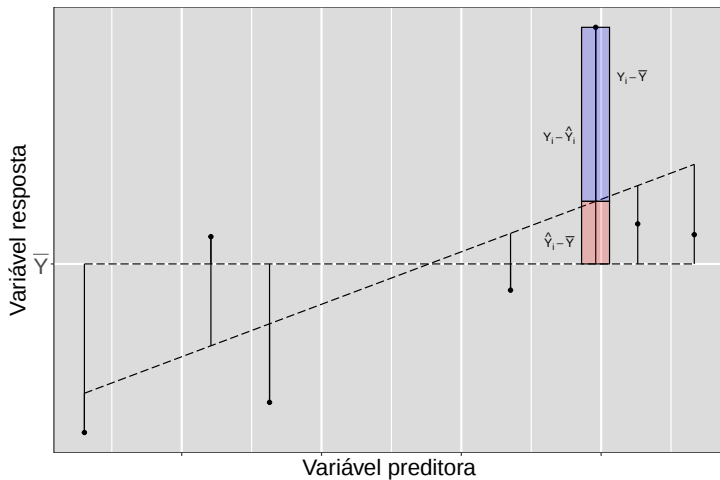


Figura 4: $Y_\ell - \hat{Y}_\ell$ e $\hat{Y}_\ell - \bar{Y}$.

Soma de quadrados

De (3) e ordenando de forma conveniente, nós temos que

$$\begin{aligned}\text{SQT} &= \sum_{\ell=1}^n (Y_{\ell} - \bar{Y})^2 = \sum_{\ell=1}^n [(\hat{Y}_{\ell} - \bar{Y}) + (Y_{\ell} - \hat{Y}_{\ell})]^2 \\&= \sum_{\ell=1}^n [(\hat{Y}_{\ell} - \bar{Y})^2 + 2(\hat{Y}_{\ell} - \bar{Y})(Y_{\ell} - \hat{Y}_{\ell}) + (Y_{\ell} - \hat{Y}_{\ell})^2] \\&= \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})^2 + \underbrace{2 \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})(Y_{\ell} - \hat{Y}_{\ell})}_0 + \sum_{\ell=1}^n (Y_{\ell} - \hat{Y}_{\ell})^2 \\&= \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})^2 + \sum_{\ell=1}^n (Y_{\ell} - \hat{Y}_{\ell})^2 \\&= \text{SQReg} + \text{SQRes}.\end{aligned}$$

Soma de quadrados

De (3) e ordenando de forma conveniente, nós temos que

$$\begin{aligned}\text{SQT} &= \sum_{\ell=1}^n (Y_{\ell} - \bar{Y})^2 = \sum_{\ell=1}^n [(\hat{Y}_{\ell} - \bar{Y}) + (Y_{\ell} - \hat{Y}_{\ell})]^2 \\&= \sum_{\ell=1}^n [(\hat{Y}_{\ell} - \bar{Y})^2 + 2(\hat{Y}_{\ell} - \bar{Y})(Y_{\ell} - \hat{Y}_{\ell}) + (Y_{\ell} - \hat{Y}_{\ell})^2] \\&= \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})^2 + \underbrace{2 \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})(Y_{\ell} - \hat{Y}_{\ell})}_0 + \sum_{\ell=1}^n (Y_{\ell} - \hat{Y}_{\ell})^2 \\&= \sum_{\ell=1}^n (\hat{Y}_{\ell} - \bar{Y})^2 + \sum_{\ell=1}^n (Y_{\ell} - \hat{Y}_{\ell})^2 \\&= \text{SQReg} + \text{SQRes}.\end{aligned}$$

Soma de quadrados

Logo, nós temos que $SQT = SQReg + SQRes$, isto é, a variabilidade total dos dados se divide em duas partes:

- a variabilidade explicada pelo modelo (SQReg);



Soma de quadrados

Logo, nós temos que $SQT = SQReg + SQRes$, isto é, a variabilidade total dos dados se divide em duas partes:

- a variabilidade explicada pelo modelo (SQReg);
- a variabilidade que não pode ser explicada pela regressão (SQRes).



Soma de quadrados

Logo, nós temos que $SQT = SQReg + SQRes$, isto é, a variabilidade total dos dados se divide em duas partes:

- a variabilidade explicada pelo modelo (SQReg);
- a variabilidade que não pode ser explicada pela regressão (SQRes).



Soma de quadrados

Em notação matricial, SQT, SQReg e SQRes são dadas, respectivamente, por

$$\begin{aligned} \text{SQT} &= \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{SQReg} &= \mathbf{Y}^\top \left(\mathbf{H} - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{SQRes} &= \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}, \end{aligned} \tag{4}$$

Soma de quadrados

Em notação matricial, SST, SQRreg e SQRres são dadas, respectivamente, por

$$\begin{aligned} \text{SST} &= \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{SQRreg} &= \mathbf{Y}^\top \left(\mathbf{H} - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{SQRres} &= \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}, \end{aligned} \tag{4}$$

em que $\mathbf{J}$ é uma matriz, $n \times n$, com todos os seus elementos iguais a 1. Para demonstrações, é conveniente escrever $\frac{1}{n} \mathbf{J} = \mathbf{1}_n (\mathbf{1}_n^\top \mathbf{1}_n)^{-1} \mathbf{1}_n^\top$, em que $\mathbf{1}_n$ é um vetor de comprimento n de 1's.



Soma de quadrados

Em notação matricial, SQT, SQReg e SQRes são dadas, respectivamente, por

$$\begin{aligned} \text{SQT} &= \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{SQReg} &= \mathbf{Y}^\top \left(\mathbf{H} - \frac{1}{n} \mathbf{J} \right) \mathbf{Y}, \\ \text{SQRes} &= \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}, \end{aligned} \tag{4}$$

em que $\mathbf{J}$ é uma matriz, $n \times n$, com todos os seus elementos iguais a 1. Para demonstrações, é conveniente escrever $\frac{1}{n} \mathbf{J} = \mathbf{1}_n (\mathbf{1}_n^\top \mathbf{1}_n)^{-1} \mathbf{1}_n^\top$, em que $\mathbf{1}_n$ é um vetor de comprimento n de 1's.



Graus de liberdade

Os graus de liberdade (GL) são valores associados às somas de quadrados. Eles são o número de informações independentes, envolvendo as observações $Y_1, Y_2, \dots, Y_n$, que são necessárias para calcular as somas de quadrados.



Graus de liberdade

Basicamente, é o tamanho da amostra menos o número de restrições. Para cada soma de quadrados, nós temos que

- SQT tem $n - 1$ GL;



Graus de liberdade

Basicamente, é o tamanho da amostra menos o número de restrições. Para cada soma de quadrados, nós temos que

- SQT tem $n - 1$ GL;
- SQReg tem $p - 1$ GL;



Graus de liberdade

Basicamente, é o tamanho da amostra menos o número de restrições. Para cada soma de quadrados, nós temos que

- SQT tem $n - 1$ GL;
- SQReg tem $p - 1$ GL;
- SQRes tem $n - p$ GL.



Graus de liberdade

Basicamente, é o tamanho da amostra menos o número de restrições. Para cada soma de quadrados, nós temos que

- SQT tem $n - 1$ GL;
- SQReg tem $p - 1$ GL;
- SQRes tem $n - p$ GL.



Quadrado médio

O quadrado médio da regressão (QMReg) e o quadrado médio do resíduo (QMRes) são definidos pela, respectiva, divisão entre as somas de quadrados e os graus de liberdades, isto é,

$$\text{QMReg} = \frac{\text{SQReg}}{p - 1} \text{ e } \text{QMRes} = \frac{\text{SQRes}}{n - p}. \quad (5)$$

Quadrado médio

O quadrado médio da regressão (QMReg) e o quadrado médio do resíduo (QMRes) são definidos pela, respectiva, divisão entre as somas de quadrados e os graus de liberdades, isto é,

$$\text{QMReg} = \frac{\text{SQReg}}{p - 1} \text{ e } \text{QMRes} = \frac{\text{SQRes}}{n - p}. \quad (5)$$



Quadrado médio

Pode-se mostrar que (Montgomery et al., 2012, p. 581),

$$\mathbb{E}(\text{QMReg}) = \sigma^2 + f(\beta_2, \beta_3, \dots, \beta_p),$$

$$\mathbb{E}(\text{QMRes}) = \sigma^2,$$

em que $f(\cdot)$ é uma função positiva de $\beta_2, \beta_3, \dots, \beta_p$, igual a 0 quando $\beta_2 = \beta_3 = \dots = \beta_p = 0$.

Quadrado médio

Pode-se mostrar que (Montgomery et al., 2012, p. 581),

$$\mathbb{E}(\text{QMReg}) = \sigma^2 + f(\beta_2, \beta_3, \dots, \beta_p),$$

$$\mathbb{E}(\text{QMRes}) = \sigma^2,$$

em que $f(\cdot)$ é uma função positiva de $\beta_2, \beta_3, \dots, \beta_p$, igual a 0 quando $\beta_2 = \beta_3 = \dots = \beta_p = 0$.



Quadrado médio

Adicionalmente, quando $\beta_2 = \beta_3 = \dots = \beta_p = 0$, nós temos que,

$$F_c = \frac{\text{QMReg}}{\text{QMRes}} \sim F(p-1, n-p).$$

Quadrado médio

Adicionalmente, quando $\beta_2 = \beta_3 = \dots = \beta_p = 0$, nós temos que,

$$F_c = \frac{\text{QMReg}}{\text{QMRes}} \sim F(p-1, n-p).$$

Anova

Tabela 1: Análise de variância.

Fonte de variação	Soma de quadrados	Graus de liberdade	Quadrado médio	F_c
Regressão	SQReg	$p - 1$	QMReg	$\frac{\text{QMReg}}{\text{QMRes}}$
Resíduos	SQRes	$n - p$	QMRes	
Total	SQT	$n - 1$		



Teste de hipóteses

Seja $\beta^* = (\beta_2, \beta_3, \dots, \beta_p)^\top$, o teste de hipóteses usando o enfoque de análise de variância é da seguinte forma:

$$\begin{cases} \mathcal{H}: \beta^* = \mathbf{0}_{p-1} \\ \mathcal{A}: \beta^* \neq \mathbf{0}_{p-1} \end{cases}.$$

Teste de hipóteses

Seja $\beta^* = (\beta_2, \beta_3, \dots, \beta_p)^\top$, o teste de hipóteses usando o enfoque de análise de variância é da seguinte forma:

$$\begin{cases} \mathcal{H}: \beta^* = \mathbf{0}_{p-1} \\ \mathcal{A}: \beta^* \neq \mathbf{0}_{p-1} \end{cases}.$$

A estatística do teste é dada por:

$$F_c = \frac{\text{QMReg}}{\text{QMRes}}.$$



Teste de hipóteses

Seja $\beta^* = (\beta_2, \beta_3, \dots, \beta_p)^\top$, o teste de hipóteses usando o enfoque de análise de variância é da seguinte forma:

$$\begin{cases} \mathcal{H}: \beta^* = \mathbf{0}_{p-1} \\ \mathcal{A}: \beta^* \neq \mathbf{0}_{p-1} \end{cases}.$$

A estatística do teste é dada por:

$$F_c = \frac{\text{QMReg}}{\text{QMRes}}.$$



Teste de hipóteses

Sob $\mathcal{H}$, $F_c \sim F(p-1, n-p)$. E assim, nós rejeitaremos $\mathcal{H}$ se $F_c > F(1-\alpha; p-1, n-p)$, isto é, a hipótese nula é rejeitada, para um dado nível nominal α , se a estatística do teste for maior que o quantil $100(1-\alpha)\%$ de uma $F(p-1, n-p)$.



Teste de hipóteses

Interpretação

Nesse caso, nós assumimos que há relação linear entre a variável resposta e pelo menos uma variável regressora, ao nível α de significância.

Teste de hipóteses

Interpretação

Nesse caso, nós assumimos que há relação linear entre a variável resposta e pelo menos uma variável regressora, ao nível α de significância.

Roteiro

- 1 Introdução
- 2 Partição de somas de quadrados
- 3 Coeficiente de determinação**
- 4 Aplicação
- 5 Estudo de simulação
- 6 Referências bibliográficas



Coeficiente de determinação

$$R^2 = \frac{SQ_{Reg}}{SQ_T} = 1 - \frac{SQ_{Res}}{SQ_T}. \quad (6)$$

Note que, $0 \leq R^2 \leq 1$.



Coeficiente de determinação

$$R^2 = \frac{SQReg}{SQT} = 1 - \frac{SQRes}{SQT}. \quad (6)$$

Note que, $0 \leq R^2 \leq 1$. E ele tende a aumentar a medida que novas covariáveis são adicionada no modelo (mesmo as não significativas).



Coeficiente de determinação

$$R^2 = \frac{SQReg}{SQT} = 1 - \frac{SQRes}{SQT}. \quad (6)$$

Note que, $0 \leq R^2 \leq 1$. E ele tende a aumentar a medida que novas covariáveis são adicionada no modelo (mesmo as não significativas).

Coeficiente de determinação

Interpretação 1

Mede a redução da incerteza de se prever a variável resposta quando nós deixamos de prevê-la por sua média amostral e passamos a prever utilizando o modelo de regressão considerado.

Interpretação 2

Coeficiente de determinação

Interpretação 1

Mede a redução da incerteza de se prever a variável resposta quando nós deixamos de prevê-la por sua média amostral e passamos a prever utilizando o modelo de regressão considerado.

Interpretação 2

Mede a proporção da variabilidade da variável resposta que é explicada pelo modelo de regressão considerado.

Coeficiente de determinação

Interpretação 1

Mede a redução da incerteza de se prever a variável resposta quando nós deixamos de prevê-la por sua média amostral e passamos a prever utilizando o modelo de regressão considerado.

Interpretação 2

Mede a proporção da variabilidade da variável resposta que é explicada pelo modelo de regressão considerado.



Observações

- $R^2 = 0$ se $\beta_2 = \beta_3 = \dots = \beta_p = 0$.

Observações

- $R^2 = 0$ se $\beta_2 = \beta_3 = \dots = \beta_p = 0$.
- $R^2 = 1$ se $Y_\ell = \hat{Y}_\ell$ para todas as observações.

Observações

- $R^2 = 0$ se $\beta_2 = \beta_3 = \dots = \beta_p = 0$.
- $R^2 = 1$ se $Y_\ell = \hat{Y}_\ell$ para todas as observações.
- R^2 é o coeficiente de determinação simples de um modelo de regressão de Y_ℓ em $\hat{Y}_\ell$.

Observações

- $R^2 = 0$ se $\beta_2 = \beta_3 = \dots = \beta_p = 0$.
- $R^2 = 1$ se $Y_\ell = \hat{Y}_\ell$ para todas as observações.
- R^2 é o coeficiente de determinação simples de um modelo de regressão de Y_ℓ em $\hat{Y}_\ell$.
- Alto R^2 não indica necessariamente que se pode fazer previsões com boa acurácia.

Observações

- $R^2 = 0$ se $\beta_2 = \beta_3 = \dots = \beta_p = 0$.
- $R^2 = 1$ se $Y_\ell = \hat{Y}_\ell$ para todas as observações.
- R^2 é o coeficiente de determinação simples de um modelo de regressão de Y_ℓ em $\hat{Y}_\ell$.
- Alto R^2 não indica necessariamente que se pode fazer previsões com boa acurácia.
- Necessariamente aumenta ou mantém-se constante com a inclusão de mais uma variável preditora.

Observações

- $R^2 = 0$ se $\beta_2 = \beta_3 = \dots = \beta_p = 0$.
- $R^2 = 1$ se $Y_\ell = \hat{Y}_\ell$ para todas as observações.
- R^2 é o coeficiente de determinação simples de um modelo de regressão de Y_ℓ em $\hat{Y}_\ell$.
- Alto R^2 não indica necessariamente que se pode fazer previsões com boa acurácia.
- Necessariamente aumenta ou mantém-se constante com a inclusão de mais uma variável preditora.

Coeficiente de determinação ajustado

$$R_a^2 = 1 - \frac{\text{SQRes}/(n - p)}{\text{SQT}/(n - 1)} = 1 - \frac{n - 1}{n - p} \frac{\text{SQRes}}{\text{SQT}}. \quad (7)$$

Penaliza modelos que adicionam covariáveis pouco significativas.

Coeficiente de determinação ajustado

$$R_a^2 = 1 - \frac{\text{SQRes}/(n - p)}{\text{SQT}/(n - 1)} = 1 - \frac{n - 1}{n - p} \frac{\text{SQRes}}{\text{SQT}}. \quad (7)$$

Penaliza modelos que adicionam covariáveis pouco significativas.

Coeficiente de correlação

$$r = \sqrt{R^2}.$$

Roteiro

- 1 Introdução
- 2 Partição de somas de quadrados
- 3 Coeficiente de determinação
- 4 Aplicação**
- 5 Estudo de simulação
- 6 Referências bibliográficas



Aplicação

Conjunto de dados *Westwood Company* (Neter et al., 1983, p. 36). Após o ajuste, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = 10 + 2x_{\ell 2},$$

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



Aplicação

Conjunto de dados *Westwood Company* (Neter et al., 1983, p. 36). Após o ajuste, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = 10 + 2x_{\ell 2},$$

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



Anova

Utilizando as expressões dadas em (4) e (5) e substituindo os resultados na Tabela 1, a tabela da Anova fica da seguinte forma:

Tabela 2: Análise de variância.

Fonte de variação	Soma de quadrados	Graus de liberdade	Quadrado médio	F_c
Regressão	13.600	1	13.600	1.813,3
Resíduos	60	8	7,5	
Total	13.660	9		

Anova

Utilizando as expressões dadas em (4) e (5) e substituindo os resultados na Tabela 1, a tabela da Anova fica da seguinte forma:

Tabela 2: Análise de variância.

Fonte de variação	Soma de quadrados	Graus de liberdade	Quadrado médio	F_c
Regressão	13.600	1	13.600	1.813,3
Resíduos	60	8	7,5	
Total	13.660	9		

Anova

Para este conjunto de dados, o teste de hipóteses usando o enfoque de análise de variância é da seguinte forma:

$$\begin{cases} \mathcal{H}: \beta_2 = 0 \\ \mathcal{A}: \beta_2 \neq 0 \end{cases}.$$

Como $F_c = 1.1813,3 > 5,3 = F(0,95; 1, 8)$,

Anova

Para este conjunto de dados, o teste de hipóteses usando o enfoque de análise de variância é da seguinte forma:

$$\begin{cases} \mathcal{H}: \beta_2 = 0 \\ \mathcal{A}: \beta_2 \neq 0 \end{cases}.$$

Como $F_c = 1.1813,3 > 5,3 = F(0,95; 1, 8)$, nós rejeitamos $\mathcal{H}$, isto é, a um nível de significância de 5%, o modelo de regressão é significativo.

Anova

Para este conjunto de dados, o teste de hipóteses usando o enfoque de análise de variância é da seguinte forma:

$$\begin{cases} \mathcal{H}: \beta_2 = 0 \\ \mathcal{A}: \beta_2 \neq 0 \end{cases}.$$

Como $F_c = 1.1813,3 > 5,3 = F(0,95; 1, 8)$, nós rejeitamos $\mathcal{H}$, isto é, a um nível de significância de 5%, o modelo de regressão é significativo.

Anova

Por (6) e (7), os coeficientes determinação, são dados por

$$R^2 = 0,996 \text{ e } R_a^2 = 0,995.$$

Anova

Por (6) e (7), os coeficientes determinação, são dados por

$$R^2 = 0,996 \text{ e } R_a^2 = 0,995.$$

A proporção da variabilidade da varável resposta que é explicada pelo o modelo de regressão é, aproximadamente, de 99%.

Anova

Por (6) e (7), os coeficientes determinação, são dados por

$$R^2 = 0,996 \text{ e } R_a^2 = 0,995.$$

A proporção da variabilidade da varável resposta que é explicada pelo o modelo de regressão é, aproximadamente, de 99%.

Roteiro

- 1 Introdução
- 2 Partição de somas de quadrados
- 3 Coeficiente de determinação
- 4 Aplicação
- 5 Estudo de simulação**
- 6 Referências bibliográficas



Estudo de simulação

Suponham o seguinte MNL,

$$Y_{\ell} = 1 + 0x_{\ell 2} + \varepsilon_{\ell},$$

em que $x_{\ell 2}$ é quantidade conhecida e $\varepsilon_{\ell} \sim \mathcal{N}(0, 4)$, $\ell = 1, 2, \dots, n$. E nós vamos assumir que $x_{\ell 2} \sim$ uniforme no intervalo $(0, 1)$. A covariável $x_{\ell 2}$, para cada n ($n = 10, 30, \dots, 110$) é mantida fixa nas réplicas de Monte Carlo. As simulações foram feitas utilizando o R (R Core Team, 2021).



Estudo de simulação

Suponham o seguinte MNL,

$$Y_{\ell} = 1 + 0x_{\ell 2} + \varepsilon_{\ell},$$

em que $x_{\ell 2}$ é quantidade conhecida e $\varepsilon_{\ell} \sim \mathcal{N}(0, 4)$, $\ell = 1, 2, \dots, n$. E nós vamos assumir que $x_{\ell 2} \sim$ uniforme no intervalo $(0, 1)$. A covariável $x_{\ell 2}$, para cada n ($n = 10, 30, \dots, 110$) é mantida fixa nas réplicas de Monte Carlo. As simulações foram feitas utilizando o R (R Core Team, 2021).



Estudo de simulação

Nós faremos um estudo de Monte Carlo, com 10.000 réplicas, onde avaliaremos

$$\begin{cases} \mathcal{H}: \beta_2 = 0 \\ \mathcal{A}: \beta_2 \neq 0 \end{cases},$$

Estudo de simulação

Nós faremos um estudo de Monte Carlo, com 10.000 réplicas, onde avaliaremos

$$\begin{cases} \mathcal{H}: \beta_2 = 0 \\ \mathcal{A}: \beta_2 \neq 0 \end{cases},$$

através do tamanho do teste,

Estudo de simulação

Nós faremos um estudo de Monte Carlo, com 10.000 réplicas, onde avaliaremos

$$\begin{cases} \mathcal{H} : \beta_2 = 0 \\ \mathcal{A} : \beta_2 \neq 0 \end{cases},$$

através do tamanho do teste, i.e.,

$$\alpha = \mathbb{P}(\text{Rejeitar } \mathcal{H} | \mathcal{H} \text{ é verdade}).$$



Estudo de simulação

Nós faremos um estudo de Monte Carlo, com 10.000 réplicas, onde avaliaremos

$$\begin{cases} \mathcal{H}: \beta_2 = 0 \\ \mathcal{A}: \beta_2 \neq 0 \end{cases},$$

através do tamanho do teste, i.e.,

$$\alpha = \mathbb{P}(\text{Rejeitar } \mathcal{H} | \mathcal{H} \text{ é verdade}).$$



Estudo de simulação

Assumindo que $\mathcal{H}$ é verdade, o tamanho empírico do teste via Anova é calculado como a proporção de vezes que a estatística do teste (F_c) foi maior que o valor F_0 nas 10.000 replicas de Monte Carlo, para $\alpha = 10\%$, 5% e 1% .



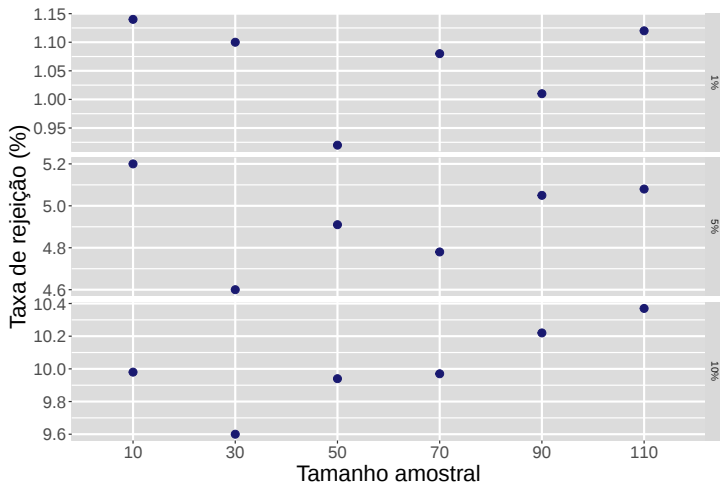


Figura 5: Tamanho do teste.

Roteiro

- 1 Introdução
- 2 Partição de somas de quadrados
- 3 Coeficiente de determinação
- 4 Aplicação
- 5 Estudo de simulação
- 6 Referências bibliográficas

Referências bibliográficas I

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.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

URL: <https://www.R-project.org/>



Obrigado!

✉ tiago.magalhaes@ufjf.br

📄 ufjf.br/tiago_magalhaes

🌐 Departamento de Estatística, Sala 319

