Modelos Lineares I

Regressão Linear Múltipla (RLM): Introdução

(17^a, 18^a, 19^a e 20^a Aulas)

16^a Aula → 1^a VE de Modelos Lineares I (*RLS:* 16/09/19)

Professor: Dr. José Rodrigo de Moraes
Universidade Federal Fluminense (UFF)
Departamento de Estatística (GET)

1

Modelo de Regressão Linear Múltipla (RLM):

- Em regressão linear múltipla (RLM) desejamos estabelecer uma relação linear entre uma variável dependente Y e (p-1) variáveis independentes X_1, X_2, ..., X_{p-1}. O modelo de regressão linear múltipla é da forma:

$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+ \ldots + \beta_{p-1} X_{p-1} + \varepsilon$

sendo chamado de modelo de 1^a ordem com (p-1) variáveis independentes.

$$Y = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_k + \varepsilon$$

$$Y = \sum_{k=0}^{p-1} \beta_k X_k + \varepsilon, \quad \text{onde}: \quad X_0 = 1.$$

2

Modelo de Regressão Linear Múltipla (RLM):

- Supondo que $E(\varepsilon)=0$, a função resposta para o modelo de regressão linear múltipla é dada por:

$E(Y)=\beta_0+\beta_1 X_1+\beta_2 X_2+ \ldots + \beta_{p-1} X_{p-1}$

Observações:

- ✓ Se p-1=1, o modelo corresponde ao modelo de RLS.
- ✓ O parâmetro β_k, k=1,2,...,p-1 é a variação em E(Y) devido ao acréscimo de 1 unidade na variável X_k.

- Para se obter estimativas para os parâmetros β_k são realizadas n observações da variável Y, ou sejam Y_i, i=1,2,..,n; conforme mostrado a seguir:

3

Modelo de Regressão Linear Múltipla (RLM):

- A variável X_k pode ser identificada por X_{ik}, onde:

X_{ik} → valor da k-ésima variável explicativa referente a i-ésima observação, i=1,2,..., n e k=1,2,..., p-1.

De um modo geral as n observações serão denotadas pelas n equações abaixo:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$

Para i=1,2,..., n; obtemos as n equações seguintes:

$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \ldots + \beta_{p-1} X_{1,p-1} + \varepsilon_1$

$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \ldots + \beta_{p-1} X_{2,p-1} + \varepsilon_2$

....

$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_{p-1} X_{n,p-1} + \varepsilon_n$

4

Modelo de Regressão Linear Múltipla (RLM):

- Uma forma simples e útil de representar o modelo de regressão linear múltipla (RLM): **Y**= Xβ + ε

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{nx1} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}_{nxp} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{px1} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{nx1}$$

- **Modelo de RLM: Y**= Xβ + ε , onde:

- β é um vetor de parâmetros desconhecidos;
- X é uma matriz de valores fixados;
- ε é um vetor aleatório com distribuição normal de:

$E(\varepsilon)=\mathbf{0}$ e $E(\varepsilon\, \varepsilon`)=\sigma^2\, \mathbf{I}_n$.

5

Modelo de Regressão Linear Múltipla (RLM):

- Com relação as hipóteses do modelo, temos que:

$$E(\boldsymbol{\varepsilon}) = E\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}`) = E\begin{bmatrix} \varepsilon_1\varepsilon_1 & \varepsilon_1\varepsilon_2 & & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2\varepsilon_2 & & \varepsilon_2\varepsilon_n \\ \vdots & \vdots & & \vdots \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & & \varepsilon_n\varepsilon_n \end{bmatrix} = \begin{bmatrix} VAR(\varepsilon_1) & COV(\varepsilon_1\varepsilon_2) & \cdots & COV(\varepsilon_1\varepsilon_n) \\ COV(\varepsilon_2\varepsilon_1) & VAR(\varepsilon_2) & \cdots & COV(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & & \vdots \\ COV(\varepsilon_n\varepsilon_1) & COV(\varepsilon_n\varepsilon_2) & \cdots & VAR(\varepsilon_n) \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n$$

6

Modelo de Regressão Linear Múltipla Normal

Exercício proposto 1: Usando a estrutura matricial **Y=Xβ+ε** represente o modelo de regressão linear simples normal (p-1=1) destacando os seus componentes.

7

Modelo de Regressão Linear Múltipla:
Estimadores de MQO do Vetor de Parâmetros β.

- Analogamente a RLS, o método dos MQO consiste em minimizar soma S dos quadrados das diferenças entre os valores observados Y_i e suas médias $E(Y_i)$, ou seja:

$Y_i = \beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+ \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$

$E(Y_i) = \beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+ \ldots + \beta_{p-1} X_{i,p-1}$

- De forma que: $\varepsilon_i = Y_i - E(Y_i)$

$\varepsilon_i = Y_i - E(Y_i) = Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \ldots - \beta_{p-1} X_{i,p-1}$

- Em termos matriciais:

$\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$

8

Modelo de Regressão Linear Múltipla:
Estimação por Mínimos Quadrados ordinários (MQO) do vetor de parâmetros β.

- A soma dos quadrados dos erros pode ser escrita matricialmente, como segue:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}`\boldsymbol{\varepsilon}$$

$$S = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})`(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

$$S = \mathbf{Y}`\mathbf{Y} - 2\cdot\boldsymbol{\beta}`\mathbf{X}`\mathbf{Y} + \boldsymbol{\beta}`(\mathbf{X}`\mathbf{X})\boldsymbol{\beta}$$

Derivando S em relação a β:

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = 0 \;\rightarrow\; \hat{\boldsymbol{\beta}} = (\mathbf{X}`\mathbf{X})^{-1}\mathbf{X}`\mathbf{Y}$$

9

Modelo de regressão linear múltipla:
Estimador de Mínimos Quadrados Ordinários (MQO) do vetor de parâmetros β.

- Desse modo, o **modelo de RLM ajustado** é dado por:

$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, onde:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}`\mathbf{X})^{-1}\mathbf{X}`\mathbf{Y}$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}`\mathbf{X})^{-1}\mathbf{X}`\boldsymbol{\varepsilon}$$

Fórmula ramificada

10

Modelo de regressão linear múltipla

Exercício proposto 2: No caso do modelo de regressão linear simples (RLS), temos que:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{nx1} \qquad \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}_{nxp} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{2x1}$$

Obtenha novamente as estimativas de β_0 e β_1, usando a expressão:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}`\mathbf{X})^{-1}\mathbf{X}`\mathbf{Y}$$

11

Modelo de regressão linear múltipla

Exercício proposto 3: No caso do modelo de regressão linear simples (RLS), mostre que:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\,\overline{X} = \frac{\overline{Y}\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

12

Prof.: José Rodrigo de Moraes: Estatístico (ENCE), Mestre em Estatística Social (ENCE) e Doutor em Saúde Coletiva (IESC/UFRJ)

Modelo de Regressão Linear Múltipla:

Cálculo da média do estimador de β:

$$E(\hat{\boldsymbol\beta}) = \boldsymbol\beta$$

OBS: O vetor de estimadores de MQO é composto por estimadores não tendenciosos dos parâmetros β_k, isto é:

$$E(\hat{\beta}_k) = \beta_k \;, \quad \forall \quad k=0,1,\ldots,p-1$$

Cálculo da variância do estimador de β:

Como $E(\hat{\beta}_k) = \beta_k$, então a variância do estimador de β_k é dado por:

$$VAR(\hat{\beta}_k) = E\left[(\hat{\beta}_k - \beta_k)^2\right], \quad \forall \quad k = 0,1,\ldots,p-1$$

13

Modelo de Regressão Linear Múltipla (*continuação*):

- **Matriz de variância-covariância:**

$$VAR(\hat{\boldsymbol\beta}) = E\left[(\hat{\boldsymbol\beta} - \boldsymbol\beta)(\hat{\boldsymbol\beta} - \boldsymbol\beta)'\right]$$

$$VAR(\hat{\boldsymbol\beta}) = E\begin{bmatrix} (\hat{\beta}_0-\beta_0)^2 & (\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1) & \cdots & (\hat{\beta}_0-\beta_0)(\hat{\beta}_{p-1}-\beta_{p-1}) \\ (\hat{\beta}_1-\beta_1)(\hat{\beta}_0-\beta_0) & (\hat{\beta}_1-\beta_1)^2 & \cdots & (\hat{\beta}_1-\beta_1)(\hat{\beta}_{p-1}-\beta_{p-1}) \\ \vdots & \vdots & & \vdots \\ (\hat{\beta}_{p-1}-\beta_{p-1})(\hat{\beta}_0-\beta_0) & (\hat{\beta}_{p-1}-\beta_{p-1})(\hat{\beta}_1-\beta_1) & \cdots & (\hat{\beta}_{p-1}-\beta_{p-1})^2 \end{bmatrix}$$

14

Modelo de Regressão Linear Múltipla (*continuação*):

- **Matriz de variância-covariância:**

$$VAR(\hat{\boldsymbol\beta}) = \begin{bmatrix} E(\hat{\beta}_0-\beta_0)^2 & E\left[(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)\right] & \cdots & E\left[(\hat{\beta}_0-\beta_0)(\hat{\beta}_{p-1}-\beta_{p-1})\right] \\ E\left[(\hat{\beta}_1-\beta_1)(\hat{\beta}_0-\beta_0)\right] & E\left[(\hat{\beta}_1-\beta_1)^2\right] & \cdots & E\left[(\hat{\beta}_1-\beta_1)(\hat{\beta}_{p-1}-\beta_{p-1})\right] \\ \vdots & \vdots & & \vdots \\ E\left[(\hat{\beta}_{p-1}-\beta_{p-1})(\hat{\beta}_0-\beta_0)\right] & E\left[(\hat{\beta}_{p-1}-\beta_{p-1})(\hat{\beta}_1-\beta_1)\right] & \cdots & E\left[(\hat{\beta}_{p-1}-\beta_{p-1})^2\right] \end{bmatrix}$$

$$VAR(\hat{\boldsymbol\beta}) = \begin{bmatrix} VAR(\hat{\beta}_0) & COV(\hat{\beta}_0,\hat{\beta}_1) & \cdots & COV(\hat{\beta}_0,\hat{\beta}_{p-1}) \\ COV(\hat{\beta}_0,\hat{\beta}_1) & VAR(\hat{\beta}_1) & \cdots & COV(\hat{\beta}_1,\hat{\beta}_{p-1}) \\ \vdots & \vdots & & \vdots \\ COV(\hat{\beta}_0,\hat{\beta}_{p-1}) & COV(\hat{\beta}_1,\hat{\beta}_{p-1}) & \cdots & VAR(\hat{\beta}_{p-1}) \end{bmatrix}$$

15

Modelo de Regressão Linear Múltipla:

Matriz de variância-covariância

Então a variância do estimador de β é calculada por:

$$VAR(\hat{\boldsymbol\beta}) = E\left[(\hat{\boldsymbol\beta}-\boldsymbol\beta)(\hat{\boldsymbol\beta}-\boldsymbol\beta)'\right] = E\left[(\mathbf{X'X})^{-1}\mathbf{X'}\varepsilon\varepsilon'\mathbf{X}(\mathbf{X'X})^{-1}\right] = \sigma^2\,(\mathbf{X'X})^{-1}$$

$$VAR(\hat{\boldsymbol\beta}) = \sigma^2\,(\mathbf{X'X})^{-1}$$

16

Modelo de Regressão Linear Múltipla:

Propriedade:

Conclui-se que:

$$\mathbf{Y} \sim N\left[\mathbf{X}\boldsymbol\beta \,,\, \sigma^2\,\mathbf{I}\right] \;\rightarrow\; \hat{\boldsymbol\beta} \sim N\left[\boldsymbol\beta \,,\, \sigma^2\,(\mathbf{X'X})^{-1}\right]$$

Como estimar σ^2 ?

17

Modelo de Regressão Linear Múltipla:

Estimador da variância σ^2:

- Resíduos do modelo (*forma matricial*):

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol\beta}$$

$$\mathbf{e} = \underbrace{\left[\mathbf{I}_n - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\right]}_{\mathbf{A}}\varepsilon$$

O resíduos e`s são uma combinação linear dos erros $\varepsilon$`s.

OBS: A matriz A= $\mathbf{I}_n$ - X(X`X)$^{-1}$X` é simétrica e idempotente:

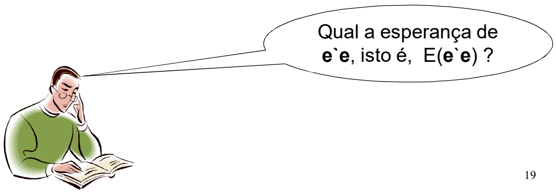
- ✓ **Simétrica: A=A`**
- ✓ **Idempotente: A^2=A**

18

Modelo de Regressão Linear Múltipla (RLM):
Estimador da variância σ^2:

- *Soma dos Quadrados dos Resíduos (SQRes):*

$$\mathbf{e}`\mathbf{e} = \varepsilon`\mathbf{A}`\mathbf{A}\,\varepsilon = \varepsilon`\mathbf{A}\,\varepsilon$$
$$\mathbf{e}`\mathbf{e} = \varepsilon`\left[\mathbf{I}_n - \mathbf{X}(\mathbf{X}`\mathbf{X})^{-1}\mathbf{X}`\right]\varepsilon$$



19

Modelo de Regressão Linear Múltipla (RLM):
Estimador da variância σ^2:

- **Propriedades importantes:**

1. Se M é uma matriz quadrada de dimensão n e se: $E(\varepsilon_i)=0$ e $VAR(\varepsilon_i)=\sigma^2\ \forall\ i=1,2,...,n$, então: $E[\varepsilon`M\,\varepsilon]=\sigma^2\,tr(M)$.

Exemplo:

$$E\left\{\begin{bmatrix}\varepsilon_1 & \varepsilon_2\end{bmatrix}\begin{bmatrix}2 & 8\\ 3 & 5\end{bmatrix}\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix}\right\} = E\left\{2\varepsilon_1^2 + 11\,\varepsilon_1\varepsilon_2 + 5\varepsilon_1^2\right\} = 7\sigma^2$$

20

Modelo de Regressão Linear Múltipla (RLM):
Estimador da variância σ^2:

- **Propriedades importantes (*continuação*):**

2. Se M é uma matriz quadrada, então tr(M)=tr(M`).

3. Dadas as matrizes quadradas A e B, se AB e BA existem, então: tr(AB)=tr(BA).

4. Dadas as matriz quadradas A, B e C, se os produtos entre elas existem, então: tr(ABC)=tr(BAC)=tr(CAB).

5. Dadas duas matrizes quadradas A e B, então: tr(A-B)=tr(A)-tr(B).

21

Modelo de Regressão Linear Múltipla (RLM):
Estimador da variância σ^2:

- Esperança da Soma dos Quadrados dos Resíduos (SQRes):

$$\mathbf{E}(\mathbf{e}`\mathbf{e}) = E\left\{\varepsilon`\left[\mathbf{I}_n - X(X`X)^{-1}X`\right]\varepsilon\right\} = \sigma^2\cdot tr\left(\mathbf{I}_n - X(X`X)^{-1}X`\right) =$$
$$= \sigma^2\cdot\left[tr(\mathbf{I}_n) - tr\left(X(X`X)^{-1}X`\right)\right] = \sigma^2\cdot\left[tr(\mathbf{I}_n) - tr\left((X`X)^{-1}(X`X)\right)\right] =$$
$$= \sigma^2\cdot\left[tr(\mathbf{I}_n) - tr(\mathbf{I}_p)\right] = \sigma^2\cdot[n-p]$$

Logo:

$$\frac{1}{n-p}\mathbf{E}(\mathbf{e}`\mathbf{e}) = \sigma^2 \;\rightarrow\; \mathbf{E}\left(\frac{\mathbf{e}`\mathbf{e}}{n-p}\right) = \sigma^2 \;\rightarrow\; \hat{\sigma}^2 = \frac{\mathbf{e}`\mathbf{e}}{n-p} \;\rightarrow\; \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

22

Inferência sobre o modelo de RLM Normal:

Já mostramos que:

$$\hat{\boldsymbol{\beta}} \sim N\left[\boldsymbol{\beta}, \sigma^2(\mathbf{X}`\mathbf{X})^{-1}\right]$$

ou seja, o vetor $\hat{\boldsymbol{\beta}}$ tem distribuição normal com:

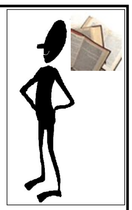
$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad e \quad \mathbf{VAR}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}`\mathbf{X})^{-1}$$

A estimativa da matriz de variância-covariância é dada por:

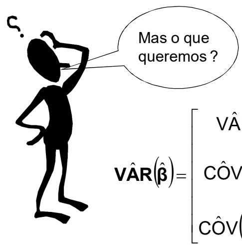
$$\mathbf{V\hat{A}R}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}`\mathbf{X})^{-1}$$

23

Modelo de RLM Normal (*continuação*):



- **Matriz de variância-covariância estimada:**



$$\mathbf{V\hat{A}R}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} V\hat{A}R(\hat{\beta}_0) & C\hat{O}V(\hat{\beta}_0,\hat{\beta}_1) & \cdots & C\hat{O}V(\hat{\beta}_0,\hat{\beta}_{p-1}) \\ C\hat{O}V(\hat{\beta}_0,\hat{\beta}_1) & V\hat{A}R(\hat{\beta}_1) & \cdots & C\hat{O}V(\hat{\beta}_1,\hat{\beta}_{p-1}) \\ \vdots & \vdots & & \vdots \\ C\hat{O}V(\hat{\beta}_0,\hat{\beta}_{p-1}) & C\hat{O}V(\hat{\beta}_1,\hat{\beta}_{p-1}) & \cdots & V\hat{A}R(\hat{\beta}_{p-1}) \end{bmatrix}$$

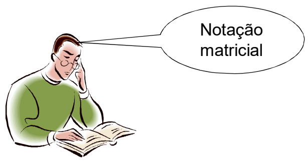
24

Prof.: José Rodrigo de Moraes: Estatístico (ENCE), Mestre em Estatística Social (ENCE) e Doutor em Saúde Coletiva (IESC/UFRJ)

Modelo de Regressão Linear Múltipla (RLM):
<u>Análise de Variância do Modelo (ANOVA):</u>

- Decomposição da variância total do modelo

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{\textbf{SQT}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\textbf{SQReg}} + \underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{\textbf{SQRes}}$$



25

Modelo de Regressão Linear Múltipla:
Análise de Variância (ANOVA) do Modelo

- <u>Componentes da ANOVA – Forma matricial de SQT:</u>

$$SQT = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = \mathbf{Y'Y} - n\bar{Y}^2$$

ou então :

$$SQT = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = Y'Y - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 = \mathbf{Y'Y} - \frac{1}{n}\mathbf{Y'UY} = \mathbf{Y'}\left[\mathbf{I}_n - \frac{1}{n}\mathbf{U}\right]\mathbf{Y}$$

onde :

U é uma matriz quadrada de dimensão n x n, tal que : $[u_{ij} = 1], \forall\, i,j = 1,2,\ldots,n$

26

Modelo de Regressão Linear Múltipla:
Análise de Variância (ANOVA)

- <u>Componentes da ANOVA – Forma matricial (*continuação*):</u>

$$SQRes = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e'e} = (\mathbf{Y} - \mathbf{X}\hat{\beta})(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

$$SQRes = \mathbf{Y'Y} - 2\cdot\hat{\beta}'\mathbf{X'Y} + \hat{\beta}'(\mathbf{X'X})\hat{\beta}$$

$$\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

$$SQRes = \mathbf{Y'Y} - \hat{\beta}'\mathbf{X'Y}$$

ou então :

$$SQRes = \mathbf{Y'}\left[\mathbf{I}_n - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\right]\mathbf{Y}$$

$$SQRes = \mathbf{Y'}\left[\mathbf{I}_n - \mathbf{H}\right]\mathbf{Y}$$, **onde : H** → matriz hat (ou de projeção)

27

Modelo de Regressão Linear Múltipla:
Análise de Variância (ANOVA)

- <u>Componentes da ANOVA – Forma matricial (*continuação*):</u>

$$SQReg = SQT - SQRes = \mathbf{Y'Y} - n\bar{Y}^2 - (\mathbf{Y'Y} - \hat{\beta}'\mathbf{X'Y}) = \hat{\beta}'\mathbf{X'Y} - n\bar{Y}^2$$

ou então :

$$SQReg = SQT - SQRes = \mathbf{Y'Y} - \frac{1}{n}\mathbf{Y'UY} - \mathbf{Y'}(\mathbf{I}_n - \mathbf{H})\mathbf{Y} =$$

$$= \mathbf{Y'HY} - \frac{1}{n}\mathbf{Y'UY} = \mathbf{Y'}\left[\mathbf{H} - \frac{1}{n}\mathbf{U}\right]\mathbf{Y}$$

onde :
U → matriz quadrada de dimensão nxn de valores unitários.
H → matriz de projeção (ou matriz Hat) dada por : $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$

28

Tabela de Análise de Variância (ANOVA) do modelo de regressão linear múltipla (RLM):

Com base nas somas dos quadrados definidos constrói-se a chamada **Tabela de análise de variância (ANOVA)**, que são utilizadas para testar a existência de relação linear entre as variáveis X`s e Y:

29

Tabela de Análise de Variância (ANOVA) - RLM:

Fontes de variação	Soma dos quadrados	gl	Quadrado médio	Estatística de teste
Regressão	$SQReg = \hat{\beta}'\mathbf{X'Y} - n\bar{Y}^2$	p-1	$QMReg = \frac{SQReg}{p-1}$	$F = \frac{QMReg}{QMRes} \sim F_{p-1,n-p}$
Resíduos	$SQRes = \mathbf{Y'Y} - \hat{\beta}'\mathbf{X'Y}$	n - p	$QMRes = \frac{SQRes}{n-p}$	
Total	$SQT = \mathbf{Y'Y} - n\bar{Y}^2$	n -1		

30

Prof.: José Rodrigo de Moraes: Estatístico (ENCE), Mestre em Estatística Social (ENCE) e Doutor em Saúde Coletiva (IESC/UFRJ)

Modelo de Regressão Linear Múltipla (RLM):
Teste de Significância Geral (tabela ANOVA)

- Para o modelo de regressão linear múltipla, a análise de variância se resume na construção do teste estatístico:

❑ **Hipóteses a serem testadas:**

$$\begin{cases} H_0: \beta_1 = \beta_1 = \cdots \beta_{p-1} = 0 \\ H_1: \beta_k \neq 0 \quad \text{para algum } k = 1,2,\ldots,p-1 \end{cases}$$

❑ **Estatística de Teste:**

$$\frac{(n-p)\text{QMRes}}{\sigma^2} \sim \chi^2_{n-p} \quad \text{e} \quad \frac{(p-1)\text{QMReg}}{\sigma^2} \sim \chi^2_{p-1}$$

31

Modelo de Regressão Linear Múltipla (RLM):
Teste de Significância Geral (tabela ANOVA)

❑ **Estatística de Teste (*continuação*):**

$$F = \frac{\frac{(p-1)\text{QMReg}}{\sigma^2} \Big/ p-1}{\frac{(n-p)\text{QMRes}}{\sigma^2} \Big/ n-p} = \frac{\text{SQReg}/p-1}{\text{SQRes}/n-p} = \frac{\text{QMReg}}{\text{QMRes}} \sim F_{p-1,n-p}$$

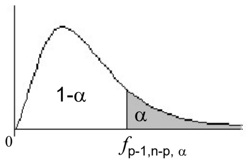
- A estatística F tem distribuição F de Snedecor com (p-1) e (n-p) graus de liberdade (F ~ $F_{p-1,n-p}$).

32

Teste de Significância Geral (tabela ANOVA) - *continuação*:

❑ **Região crítica:**

$$RC = \{ f \in \Re \ / \ f \geq f_{p-1,n-p,\alpha} \}$$



❑ **Tomada de Decisão:**

- Se $f_{obs} \in$ RC rejeita-se H_0:β_1=0 ao nível de significância α, e conclui-se que existe relação linear estatisticamente significante entre Y e pelo menos uma variável explicativa.
- Se $f_{obs} \notin$ RC não há evidências para rejeitar H_0:β_1=0 ao nível de significância α, e conclui-se que não existe relação linear estatisticamente significante entre Y e as vars explicativas do modelo.

OBS: *Pode-se utilizar também a abordagem do p-valor !!!*

33

Exemplo de aplicação: Modelo de RLM com p-1=2 variáveis explicativas

A tabela a seguir fornece o valor dos salários (em 100 UM), a idade (em anos) e o tempo de serviço (em anos) de n=25 funcionários de uma pequena empresa

O objetivo do estudo é estudar a relação entre Y e as seguintes variáveis explicativas:

✓ Idade (X_1)

✓ Tempo de serviço (X_2)

34

Tabela 1: ***Dados sobre n=25 funcionários de uma empresa***

continuação

Func.	Salário	Idade	Tempo de serviço	Func.	Salário	Idade	Tempo de serviço
1	35	48	15	16	17	21	1
2	25	25	2	17	29	45	21
3	22	23	1	18	27	40	17
4	39	55	20	19	35	43	20
5	23	40	8	20	19	23	5
6	30	42	10	21	25	30	10
7	26	24	4	22	29	31	13
8	30	38	6	23	32	35	17
9	38	49	19	24	28	34	15
10	40	52	22	25	19	21	3
11	45	57	25				
12	37	47	17				
13	43	48	25				
14	22	22	1				
15	27	48	7				

35

Modelo de RLM com p-1=2 variáveis explicativas:

Exemplo (*continuação*):

Pergunta: Pelo menos uma das variáveis tem efeito sobre o salários dos funcionários ?

Modelo completo – Representação:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

36

Modelo de RLM Normal:

Modelo teórico: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

Defina:

$Y_i \rightarrow$ ______________________________

$X_{i1} \rightarrow$ ______________________________

$X_{i2} \rightarrow$ ______________________________

$\beta_0 \rightarrow$ ______________________________

$\beta_1 \rightarrow$ ______________________________

$\beta_2 \rightarrow$ ______________________________

$\varepsilon_i \rightarrow$ ______________________________

37

Tabela de Análise da Variância (ANOVA) – RLM Normal:

Analyse /Regression / Linear

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1179,505	2	589,752	54,530	,000[a]
	Residual	237,935	22	10,815		
	Total	1417,440	24			

a. Predictors: (Constant), Tempo_serv, Idade

b. Dependent Variable: Salário_Y

Soma dos quadrados (*SQ*) e quadrados médios (*QM*).

38

Tabela de Análise da Variância (ANOVA) – RLM Normal:

Analyse /Regression / Linear

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1179,505	2	589,752	54,530	,000[a]
	Residual	237,935	22	10,815		
	Total	1417,440	24			

a. Predictors: (Constant), Tempo_serv, Idade

b. Dependent Variable: Salário_Y

TH com base na Tabela ANOVA, ao nível de 5%

39

Coeficiente de Correlação Múltipla

- O coeficiente de correlação múltipla, denotado por R, é dado por:

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

Intervalo de variação : $0 \le R \le 1$

40

Coeficiente de Múltipla Correlação (*R*): RLM Normal

Analyse /Regression / Linear

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,912[a]	,832	,817	3,289

a. Predictors: (Constant), Tempo_serv, Idade

b. Dependent Variable: Salário_Y

Coeficiente de múltipla correlação (*R*=0,912)

41

Medida de Qualidade do Ajuste:
Coeficiente de Determinação Múltipla

- O coeficiente de determinação múltipla, denotado por R^2, é dado por:

$$R^2 = \frac{SQReg}{SQT} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2}{\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2} \quad \text{ou} \quad R^2 = 1 - \frac{SQRes}{SQT} = 1 - \left[\frac{\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}}{\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2}\right]$$

- O R^2 representa a proporção (%) da variação total dos valores da variável resposta Y que é explicada pelo modelo. **Intervalo de variação:** $0\% \le R^2 \le 100\%$.
  - Se $R^2=0 \rightarrow \beta_k=0, \forall\, k=1, 2, ..., p-1$
  - Se $R^2=1 \rightarrow$ relação linear perfeita !!!

42

Coeficiente de determinação múltipla do modelo (R^2): medida de qualidade do ajuste

Analyse /Regression / Linear

Qual a interpretação do R^2 no contexto do problema ?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,912[a]	,832	,817	3,289

a. Predictors: (Constant), Tempo_serv, Idade

b. Dependent Variable: Salário_Y

Coeficiente de determinação múltipla (R^2=83,2%)

43

Medida de qualidade do ajuste:

Coeficiente de determinação múltipla

- Vimos que o R^2 é uma medida de avaliação da adequação do modelo ajustado a ser utilizado para representar os dados observados.
- Entretanto, o R^2 sempre aumenta a medida que aumenta o número de variáveis (ou parâmetros) no modelo. Devido a esta propriedade do R^2, este não deve ser utilizado para comparar 2 ou mais modelos com números distintos de variáveis (ou parâmetros). Caso contrário, o modelo com mais variáveis sempre levará uma vantagem inicial. Para contornar este problema, utiliza-se uma medida de qualidade do ajuste mais refinada, conhecida por R^2 ajustado.

44

Medida de qualidade do ajuste:

Coeficiente de determinação múltipla Ajustado

- O coeficiente de determinação múltipla ajustado, denotado por R^2_{ajust}:

$$R^2_{ajust} = 1 - \frac{SQRes/(n-p)}{SQT/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SQRes}{SQT}$$

Pesquisar: Qual a relação matemática entre R^2_{ajust} e R^2 ?

45

Coeficiente de determinação múltipla ajustado (R^2_{ajust}): medida de qualidade do ajuste

Analyse /Regression / Linear

Qual a interpretação do R^2_{ajust} no contexto do problema ?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,912[a]	,832	,817	3,289

a. Predictors: (Constant), Tempo_serv, Idade

b. Dependent Variable: Salário_Y

Coeficiente de determinação múltipla ajustado (R^2_{ajust} = 81,7%)

46