Modelos Lineares I

Regressão Linear Múltipla (RLM):

Variáveis indicadoras (*dummy*)

(35ª e 36ª Aulas)

Professor: Dr. José Rodrigo de Moraes

Universidade Federal Fluminense (UFF)

Departamento de Estatística (GET)

1

Variável indicadora (ou *dummy*):

Consideremos agora o caso no qual existe alguma variável explicativa qualitativa de interesse que se deseja incluir na modelagem estatística. Variável desse tipo dão origem a variáveis indicadoras, também chamadas de *variáveis binárias* ou *variáveis dummy*.

Os modelos de regressão, cuja matriz X é composta por variáveis indicadoras (ou *variáveis dummy*), também podem ser representados por:

2

Modelo de regressão linear múltipla (RLM) com variável indicadora:

Introdução:

Representação geral:

$Y = X\beta + \varepsilon$, tal que: $\varepsilon \sim N(0, \sigma^2 I_n)$

- Y e ε são vetores aleatórios de dimensão n;
- X é uma matriz de valores constantes de dimensão n x p;
- β é um vetor de dimensão p x 1 de parâmetros a serem estimados;
- I_n é uma matriz unitária de dimensão n.

3

Modelo de RLM com variável indicadora:

- **Conceitos Básicos:**

Variáveis quantitativas (ou *numéricas*): são aquelas variáveis que assumem valores expressos em números.

Variáveis qualitativas (ou *categóricas*): são aquelas variáveis que assumem valores expressos por categorias / atributos.

4

Modelo de RLM com variável indicadora:

- **Conceitos Básicos:**

Exemplos de variáveis qualitativas (ou *categóricas*):

Ex.1:

Sexo → variável qualitativa

Níveis da variável: masculino, feminino

Ex.2:

Autoavaliação de saúde → variável qualitativa

Níveis da variável: bom, regular, ruim

5

Modelo de RLM com variável indicadora:

- **Exemplos de associações:**

Exemplo 1:

Variável resposta → Salário (em R$).

Variável explicativa (categórica) → Nível de escolaridade, sexo, faixa-etária, cargo, etc.

Exemplo 2:

Variável resposta → IMC=peso/altura^2 (em kg/m^2).

Variável explicativa (categórica) → Sexo, faixa-etária, prática de atividade física, ingestão de energia adequada, etc.

6

Modelo de RLM com variável indicadora:

- **Variáveis indicadoras (ou *dummy*):**

Um método tradicional para discriminar os diferentes níveis de uma variável qualitativa consiste no uso de variáveis indicadoras (ou dummies).

Ilustração 1:

- Variável qualitativa: Sexo (masculino, feminino)
- 1 variável indicadora (ou *dummy*):

$$X_1 = \begin{cases} 1, & \text{se o indivíduo é do sexo masculino} \\ 0, & \text{se o indivíduo é do sexo feminino} \end{cases}$$

7

- **Variáveis indicadoras (ou dummy):**

Ilustração 2:

- Variável qualitativa: Autoavaliação de saúde (bom, regular, ruim)
- 2 variáveis indicadoras (ou *dummy*):

$$X_1 = \begin{cases} 1, & \text{se o indivíduo autoavaliar seu estado de saúde como "bom"} \\ 0, & \text{se o indivíduo não autoavaliar seu estado de saúde como "bom"} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{se o indivíduo autoavaliar seu estado de saúde como "regular"} \\ 0, & \text{se o indivíduo não autoavaliar seu estado de saúde como "regular"} \end{cases}$$

8

- **Variáveis indicadoras (ou dummy):**

Ilustração 2 (continuação):

2 variáveis indicadoras (ou *dummy*): X_1 e X_2

X_1	X_2	Autoavaliação de saúde (X)
1	0	Bom
0	1	Regular
0	0	Ruim

Uma variável qualitativa com k=3 níveis, deve ser representada por k-1=3-1=2 variáveis indicadoras (*dummy*), cada uma assumindo o valor 0 ou 1.

9

Exemplo 1: Modelo de RLM com variável indicadora

A coordenação de um curso de graduação realizou um estudo para avaliar se a idade do aluno (X_1) e o método de ensino (X_2) do professor influenciam no desempenho dos alunos (Y) numa determinada disciplina obrigatória.

A Tabela 1 fornece essas informações para uma amostra de n=20 alunos matriculados na disciplina.

10

Exemplo 1 – Tabela 1: ***Dados sobre n=20 alunos de graduação***

Aluno	Desempenho	Idade	Método de ensino	Método de ensino
1	4,544	22,5	expositiva+pratica	0
2	4,203	20,0	expositiva+pratica	0
3	5,010	25,0	expositiva+pratica	0
4	4,875	24,5	expositiva+pratica	0
5	4,792	23,5	expositiva+pratica	0
6	4,779	23,7	expositiva+pratica	0
7	5,226	26,5	expositiva+pratica	0
8	5,052	25,9	expositiva+pratica	0
9	4,558	22,1	expositiva+pratica	0
10	4,478	21,8	expositiva+pratica	0
11	3,350	22,4	expositiva	1
12	3,123	21,2	expositiva	1
13	3,752	24,8	expositiva	1
14	3,713	26,0	expositiva	1
15	3,470	24,3	expositiva	1
16	3,392	23,8	expositiva	1
17	3,213	22,4	expositiva	1
18	3,547	25,1	expositiva	1
19	3,349	23,2	expositiva	1
20	3,229	21,6	expositiva	1

11

- **Exemplo 1 (*continuação*):**

a) Construa o gráfico de dispersão entre a idade do aluno e seu desempenho na disciplina, considerando o método de ensino do professor.

b) Especifique e ajuste um modelo estatístico para avaliar o efeito da idade do aluno sobre o seu desempenho na disciplina. Calcule o coeficiente de determinação do modelo.

c) Especifique e ajuste um modelo estatístico para avaliar o efeito da idade sobre o desempenho, levando em conta o método de ensino do professor da disciplina. Calcule o coeficiente de determinação do modelo.

12

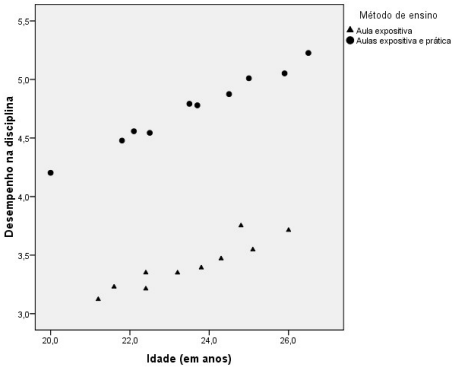
❑ **Exemplo 1 (*continuação*):**

d) Avalie a necessidade de se considerar o método de ensino na associação de interesse entre a idade e o desempenho. Para tanto, utilize o teste F de comparabilidade de modelos (construa a tabela ANOVA). Caso seja importante a inclusão do método de ensino, verifique se houve ou não mudança no efeito da idade do aluno no seu desempenho.

e) Avalie a normalidade dos resíduos do modelo escolhido usando o *QQ-Plot* ou algum teste de normalidade (*Teste Kolmogorov-Smirnov ou Shapiro-Wilk*). Qual a sua conclusão?

13

Exemplo 1 – a) Gráfico de dispersão entre a idade do aluno e seu desempenho na disciplina, considerando o método de ensino.



Exemplo 1 – b) :

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

Defina os componentes do modelo (*contexto*):

Y_i: _____

X_{i1}: _____

β_0: _____

β_1: _____

ε_i: _____

15

Exemplo 1 – b):

Resultados do ajuste do modelo com idade (X_1)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,306	1	1,306	2,644	,121[a]
	Residual	8,892	18	,494		
	Total	10,199	19			

a. Predictors: (Constant), Idade_aluno

b. Dependent Variable: Desempenho_Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	,578	2,161		,267	,792
	Idade_aluno	,149	,092	,358	1,626	,121

a. Dependent Variable: Desempenho_Y

16

Exemplo 1 – b):

Resultados do ajuste do modelo com idade (X_1)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,358[a]	,128	,080	,702869

a. Predictors: (Constant), Idade_aluno

17

Exemplo 1 – c) :

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

Defina os componentes do modelo (*contexto*):

Y_i: _____

X_{i1}: _____

X_{i2}: _____

_____

β_0: _____

β_1: _____

β_2: _____

ε_i: _____

18

Exemplo 1 – c):

Resultados do ajuste do modelo com idade (X_1), levando em conta o método de ensino (X_2)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10,121	2	5,060	1103,695	,000[a]
	Residual	,078	17	,005		
	Total	10,199	19			

a. Predictors: (Constant), Metod_ens_D, Idade_aluno

b. Dependent Variable: Desempenho_Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1,428	,209		6,827	,000
	Idade_aluno	,141	,009	,339	15,979	,000
	Metod_ens_D	-1,328	,030	-,930	-43,847	,000

a. Dependent Variable: Desempenho_Y

19

Exemplo 1 – c):

Resultados do ajuste do modelo com idade (X_1), levando em conta o método de ensino (X_2)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,996[a]	,992	,991	,067712

a. Predictors: (Constant), Metod_ens_D, Idade_aluno

20

Exemplo 1 – d):

Teste de comparabilidade de modelos (Teste F parcial: adição de 1 variável)

- **Hipóteses a serem testadas:**
  - Modelo reduzido de RLS de 1 var explicativa (idade):
  - Modelo completo de RLM de 2 vars explicativas (idade e método):

21

Exemplo 1 – d):

Teste de comparabilidade de modelos (Teste F parcial: adição de 1 variável)

- **Estatística de Teste:**

22

Exemplo 1 – d):

Teste de comparabilidade de modelos (Teste F parcial: adição de 1 variável)

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

  Cálculo do valor observado de F:

23

Exemplo 1 – d):

Teste de comparabilidade de modelos (Teste F parcial: adição de 1 variável)

- **Região crítica**:
- **Tomada de decisão:**

24

Exemplo 1 – d) Tabela ANOVA com a decomposição da SQReg em somas de quadrados extras:

Tabela 1: Para os dados dos desempenhos dos n=20 alunos, no caso de um modelo com 2 variáveis explicativas (X_1 e X_2):

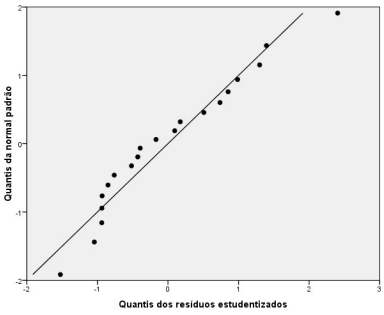
Fonte de variação	Soma dos quadrados	gl	Quadrado médio
Regressão	**SQReg(X_1,X_2)=10,121**	**2**	**QMReg(X_1,X_2)=5,060**
X_1	SQReg(X_1)=1,306	1	QMReg(X_1)=1,306
X_2/X_1	SQReg(X_2/X_1) = 8,814	1	QMReg(X_2/X_1) = 8,814
Resíduos	**SQRes(X_1,X_2) = 0,078**	**n-3=17**	**QMRes(X_1,X_2) = 0,0045882**
Total	**SQT (X_1, X_2)=10,199**	**n-1=19**	

$$f_{obs} = \frac{QMReg(X_2/X_1)}{QMRes(X_1, X_2)} = \frac{8{,}814/1}{0{,}078/17} = \frac{8{,}814}{0{,}0045882} \cong 1.921$$

25

Exemplo 1- e) Análise dos resíduos do modelo

Figura 1: ***QQ-Plot*** **(normalidade) para os resíduos estudentizados do modelo**



Teste de Kolmogorov-Smirnov: p-valor=0,767

26

Aula prática – Exercício 1 (“*Saídas*”):

Exercício 1: Estudos efetuados com recém-nascidos sugerem que características sensitivas de bebês se desenvolvem em ritmos diferentes. Enquanto, a visão de formas bem definidas ocorre apenas a partir das 4-8 semanas de vida, o olfato encontra-se completamente desenvolvido ao fim da 1ª semana de vida. Para avaliar o efeito do odor materno no tempo de adormecimento de bebês, um pesquisador efetuou um estudo no qual mediu o tempo que bebês de 1 semana, de três maternidades diferentes (A→1, B→2 e C→3), demoram a adormecer.

27

Aula prática – Exercício 1 (*continuação*):

O pesquisador em seu estudo, considerou ainda dois grupos de bebês, onde em um grupo (Grupo 1- Sim) ele colocou no berço uma peça de roupa usada pela mãe e no outro (Grupo 2 - Não), foi colocado no berço uma peça de roupa usada por outra parturiente que não a mãe.

Os dados obtidos são apresentados na tabela 2 a seguir:

28

Tabela 2: ***Dados sobre n=30 bebês de três diferentes maternidades***

bebê	Maternidade	Peça de roupa da mãe	Tempo para adormecer (min)
1	A	Sim	2
2	A	Sim	5
3	A	Sim	4
4	A	Sim	6
5	A	Sim	5
6	A	Não	9
7	A	Não	7
8	A	Não	5
9	A	Não	6
10	A	Não	5
11	B	Sim	3
12	B	Sim	6
13	B	Sim	6
14	B	Sim	5
15	B	Sim	5
16	B	Não	9
17	B	Não	7
18	B	Não	5
19	B	Não	6
20	B	Não	5

29

Tabela 2 [continuação]: ***Dados sobre n=30 bebês de três diferentes maternidades***

(continuação)

bebê	Maternidade	Peça de roupa da mãe	Tempo para adormecer (min)
21	C	Sim	9
22	C	Sim	7
23	C	Sim	7
24	C	Sim	4
25	C	Sim	5
26	C	Não	9
27	C	Não	9
28	C	Não	8
29	C	Não	9
30	C	Não	7

30

Aula prática - Exercício 1 (*continuação*):

Considerando os dados da tabela 2, pede-se:

a) Construa um gráfico para representar a relação das variáveis consideradas no estudo. Analise-o.

b) Proponha um modelo a ser ajustado aos dados observados e represente a sua equação descrevendo os termos e variáveis do modelo no contexto do problema.

c) Mostre todas as etapas de teste até a escolha do modelo final. Para tanto, utilize o Teste F de comparabilidade de modelos (Teste F parcial). **OBS:** *É preciso escrever a equação de todos os modelos sob comparação, e definir as hipóteses a serem testadas, a Estatística de teste, a Região Crítica e a Tomada de decisão.*

31

Aula prática - Exercício 1 (*continuação*):

d) Com base no modelo que você selecionou, avalie se a "peça de roupa da mãe" (odor materno) e o "tipo de maternidade" influenciam no tempo em que os bebês demoram a adormecer. Interprete os resultados do ajuste do modelo (estimativas pontuais, teste de significância individual, etc.).

e) Calcule uma medida global de qualidade do ajuste para o modelo final (interprete-a) e represente graficamente os tempos estimados para adormecer.

f) Avalie as hipóteses de normalidade e de homocedasticidade dos erros usando a análise gráfica dos resíduos estudentizados.

32

Saídas – Modelo 1:

Analyze / Regression / Linear

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	47,567	3	15,856	7,283	,001[a]
	Residual	56,600	26	2,177		
	Total	104,167	29			

a. Predictors: (Constant), Peça_roupa_G1, Maternidade_2, Maternidade_1

b. Dependent Variable: Tempo para adormecer (em minutos)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	8,300	,539		15,406	,000	7,193	9,407
	Maternidade_1	-2,000	,660	-,506	-3,031	,005	-3,356	-,644
	Maternidade_2	-1,700	,660	-,430	-2,576	,016	-3,056	-,344
	Peça_roupa_G1	-1,800	,539	-,483	-3,341	,003	-2,907	-,693

a. Dependent Variable: Tempo para adormecer (em minutos)

33

Saídas – Modelo 1:

Analyze / Regression / Linear

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,676[a]	,457	,394	1,475

a. Predictors: (Constant), Peça_roupa_G1, Maternidade_2, Maternidade_1

34

Saídas – Modelo 2:

Analyze / Regression / Linear

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23,267	2	11,633	3,883	,033[a]
	Residual	80,900	27	2,996		
	Total	104,167	29			

a. Predictors: (Constant), Maternidade_2, Maternidade_1

b. Dependent Variable: Tempo para adormecer (em minutos)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	7,400	,547		13,519	,000	6,277	8,523
	Maternidade_1	-2,000	,774	-,506	-2,584	,016	-3,588	-,412
	Maternidade_2	-1,700	,774	-,430	-2,196	,037	-3,288	-,112

a. Dependent Variable: Tempo para adormecer (em minutos)

35

Saídas – Modelo 2:

Analyze / Regression / Linear

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,473[a]	,223	,166	1,731

a. Predictors: (Constant), Maternidade_2, Maternidade_1

36

Saídas – Modelo 3:

Analyze / Regression / Linear

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24,300	1	24,300	8,519	,007[a]
	Residual	79,867	28	2,852		
	Total	104,167	29			

a. Predictors: (Constant), Peça_roupa_G1

b. Dependent Variable: Tempo para adormecer (em minutos)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	7,067	,436		16,205	,000	6,173	7,960
	Peça_roupa_G1	-1,800	,617	-,483	-2,919	,007	-3,063	-,537

a. Dependent Variable: Tempo para adormecer (em minutos)

37

Saídas – Modelo 3:

Analyze / Regression / Linear

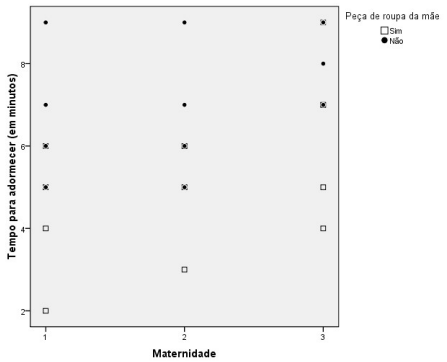
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,483[a]	,233	,206	1,689

a. Predictors: (Constant), Peça_roupa_G1

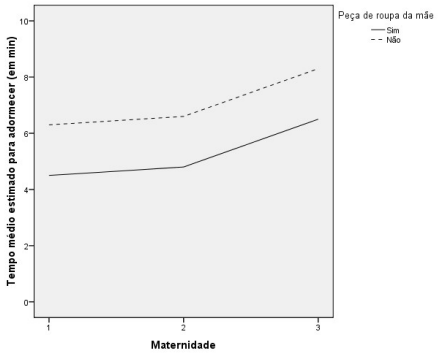
38

Figura 1: Análise gráfica dos tempos para adormecer (em min)



39

Figura 2: Análise gráfica dos tempos estimados para adormecer (em min)



40

Aula prática – Exercício 2:

Exercício 2: Considere os dados de duração de aleitamento materno exclusivo (em dias) para n=36 mães, classificadas segundo a sua escolaridade e classe social.

Faça um gráfico das durações médias de AME obtidas para cada combinação de tratamentos (escolaridade e classe social). Analise-o.

Usando os dados da tabela 3, a seguir, responda o itens seguintes. *Fazer usando o programa R e SPSS !!! É necessário descrever os testes.*

41

Tabela 3: Duração do aleitamento materno exclusivo (AME) segundo características das mães.

		Escolaridade materna		
		1-Superior	2-Médio	3-Fundamental
Classe Social	1-Baixa	130	34	20
		74	80	82
		155	40	70
		180	75	58
	2-Média	150	136	25
		159	106	58
		188	122	70
		126	115	45
	3-Alta	138	174	96
		168	150	82
		110	120	104
		160	139	60

42

Aula prática - Exercício 2 (*continuação*):

a) Proponha um modelo a ser ajustado aos dados observados e represente a sua equação descrevendo os termos e variáveis do modelo no contexto do problema.

b) Mostre todas as etapas de teste até a escolha do modelo final. Para tanto, utilize o teste F de comparabilidade de modelos (*Teste F parcial*). **OBS:** *É preciso escrever a equação de todos os modelos sob comparação; e definir as hipóteses a serem testadas, a Estatística de teste, Região Crítica e Tomada de decisão*. Interprete os resultados do ajuste do modelo (estimativas pontuais, teste de significância individual, etc.). Justifique ainda a sua escolha usando medidas de qualidade do ajuste (R^2 *ajustado*).

43

Aula prática – Exercício 2 (*continuação*):

Ainda considerando os dados da tabela 3, pede-se também:

c) Calcule uma medida global de qualidade do ajuste (R^2) para o modelo final (selecionado) e interprete-a no contexto do problema.

d) Avalie a hipótese de normalidade dos erros usando os resíduos estudentizados.

44