Modelos Lineares I

Regressão Linear Múltipla (RLM):

Seleção de modelo

(23ª, 24ª e 25ª Aulas)

Professor: Dr. José Rodrigo de Moraes

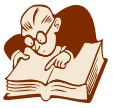
Universidade Federal Fluminense (UFF)

Departamento de Estatística (GET)

1

Modelo de Regressão Linear Múltipla Normal:

Introdução:



<u>Seleção de Modelo</u>:

Muitas das aplicações de regressão ("Modelagem Estatística") envolve um grande número de variáveis explicativas e o objetivo é identificar um subconjunto de variáveis que produz um modelo parcimonioso.

2

Modelo de Regressão Linear Múltipla (RLM):

Teste de Hipóteses em Regressão Linear Múltipla

<u>Após o ajuste do modelo</u>, deve-se considerar algumas questões com respeito ao ajuste e sobre a contribuição de cada variável explicativa para a predição de Y.

<u>Questões importantes</u>:

✓ **1) Teste sobre a contribuição global de todas as variáveis** → tratadas coletivamente, o conjunto completo das variáveis (ou, equivalentemente, o modelo ajustado) contribui significativamente para a predição de Y ?

3

Modelo de Regressão Linear Múltipla (RLM) Normal:

Teste de Hipóteses em Regressão Linear Múltipla

<u>Questões importantes (*continuação*)</u>:

✓ **2) Teste da adição de uma variável:** a adição de uma variável independente em particular melhora significativamente a predição de Y (a predição que foi alcançada pelas variáveis já existentes no modelo) ?

✓ **3) Teste de adição de um conjunto de variáveis:** a adição de um grupo de variáveis independentes melhora significativamente a predição de Y obtida pelas outras variáveis já previamente incluídas no modelo ?

4

Modelo de Regressão Linear Múltipla (RLM):

<u>Teste de comparabilidade de modelos (Teste F parcial):</u>

➢ Vimos anteriormente: Se desejamos testar se a variável X_k pode ser excluída do modelo de RLM, fixamos as hipóteses H_0: $\beta_k=0$ *contra* H_1:$\beta_k \neq 0$ e utilizamos a estatística T-Student com (n-p) g.l`s (Teste de significância individual).

➢ Para fins de seleção de modelo, pode-se usar também o **conceito de "soma dos quadrados extra"**, através do teste F parcial (teste de comparabilidade de modelos encaixados), descrito a seguir:

5

Teste de comparabilidade de modelos encaixados (Teste F parcial):

1) Hipóteses a serem testadas:

$$\begin{cases} \mathrm{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{bmatrix} & \text{Modelo reduzido} \\ & ; \quad \text{onde : } \boxed{q < p < n} \\ \mathrm{H}_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} & \text{Modelo completo} \end{cases}$$

6

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

Abordagem considerando o modelo de RLM normal:

Teste de comparabilidade de modelos (Teste F parcial):

2) Estatística de teste: Pode-se testar H_0 contra H_1 usando <u>informação</u> sobre as somas dos quadrados dos resíduos (SQRes) de ambos os modelos:

$$F = \frac{(SQRes_0 - SQRes_1)/(p-q)}{SQRes_1/(n-p)} \sim F_{p-q\,,\,n-p}$$

$SQRes_0 \rightarrow$ soma dos quadrados dos resíduos do modelo reduzido

$SQRes_1 \rightarrow$ soma dos quadrados dos resíduos do modelo completo

.

7

Modelo de Regressão Linear Múltipla (RLM) Normal:

<u>Idéia básica do "Teste de comparabilidade de modelos (Teste F parcial)":</u>

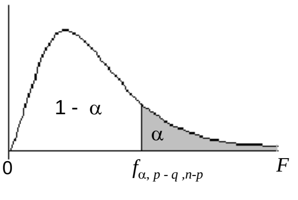
- A idéia básica do teste é verificar a redução na soma de quadrados dos resíduos (*SQRes*) quando uma ou mais variáveis explicativas são adicionadas ao modelo de regressão, dado que outras variáveis explicativas já estão incluídas no modelo.
- Por outro lado, podemos pensar no acréscimo na soma de quadrados do modelo (*SQM*) quando uma ou mais variáveis explicativas são adicionadas ao modelo.

8

Abordagem usando o modelo de RLM normal:

Teste de comparabilidade de modelos (Teste F parcial):

3) Região Crítica:



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

9

Abordagem usando o modelo de RLM normal:

Teste de comparabilidade de modelos (Teste F parcial):

4) Tomada de decisão:

- Se $f_{obs} \in \Re C$ rejeita-se H_0 ao nível de significância α, e conclui-se que o modelo reduzido <u>não é tão</u> adequado quanto o modelo completo.
- Se $f_{obs} \notin \Re C$ não há evidências para rejeitar H_0 ao nível de significância α, e conclui-se que o modelo reduzido <u>é tão</u> adequado quanto o modelo completo.

<u>Ou usando o método do p-valor:</u>

Excel: ***p-valor = DISTF (f_{obs}, gl_N, gl_D)***

Programa R: ***p-valor=1-pf(f_{obs}, gl_N, gl_D)***

10

Exemplo: Modelo de RLM com p-1=3 vars explicativas

Os dados apresentados na tabela a seguir se referem a um estudo sobre a quantidade de gordura no corpo (Y) realizado com uma amostra de n=20 mulheres saudáveis de 25 a 34 anos de idade.

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

✓ espessura da pele (X_1)

✓ perímetro da coxa (X_2)

✓perímetro braquial (X_3).

11

Banco de Dados: ***Modelo de RLM Normal (n=20 mulheres)*****:**

Mulher	Espessura da pele (X_1)	Perímetro da coxa (X_2)	Perímetro do braço (X_3)	Quantidade de gordura (Y)
1	19,5	43,1	29,1	11,9
2	24,7	49,8	28,2	22,8
3	30,7	51,9	37,0	18,7
4	29,8	54,3	31,1	20,1
5	19,1	42,2	30,9	12,9
6	25,6	53,9	23,7	21,7
7	31,4	58,5	27,6	27,1
8	27,9	52,1	30,6	25,4
9	22,1	49,9	23,2	21,3
10	25,5	53,5	24,8	19,3
11	31,1	56,6	30,0	25,4
12	30,4	56,7	28,3	27,2
13	18,7	46,5	23,0	11,7
14	19,7	44,2	28,6	17,8
15	14,6	42,7	21,3	12,8
16	29,5	54,4	30,1	23,9
17	27,7	55,3	25,7	22,6
18	30,2	58,6	24,6	25,4
19	22,7	48,2	27,1	14,8
20	25,2	51,0	27,5	21,1

12

Exemplo: Teste de comparabilidade de modelos (Teste F parcial)

Ajustes de 4 modelos de regressão linear:

Modelo 1: Regressão da quantidade de gordura (Y) sobre espessura da pele (X_1);

Modelo 2: Regressão da quantidade de gordura (Y) sobre o perímetro da coxa (X_2);

Modelo 3: Regressão da quantidade de gordura (Y) sobre espessura da pele (X_1) e sobre o perímetro da coxa (X_2);

Modelo 4: Regressão da quantidade de gordura (Y) sobre espessura da pele (X_1), o perímetro da coxa (X_2) e perímetro braquial (X_3).

13

Exemplo: Resultados do ajuste do modelo 1

Modelo 1: Regressão da quantidade de gordura (Y) sobre espessura da pele (X_1);

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	352,270	1	352,270	44,305	,000[a]
	Residual	143,120	18	7,951		
	Total	495,389	19			

a. Predictors: (Constant), Esp_pele_X1

b. Dependent Variable: Qde_gordura_Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1,496	3,319		-,451	,658
	Esp_pele_X1	,857	,129	,843	6,656	,000

a. Dependent Variable: Qde_gordura_Y

14

Exemplo: Resultados do ajuste do modelo 2

Modelo 2: Regressão da quantidade de gordura (Y) sobre o perímetro da coxa (X_2);

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	381,966	1	381,966	60,617	,000[a]
	Residual	113,424	18	6,301		
	Total	495,389	19			

a. Predictors: (Constant), Perim_coxa_X2

b. Dependent Variable: Qde_gordura_Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-23,634	5,657		-4,178	,001
	Perim_coxa_X2	,857	,110	,878	7,786	,000

a. Dependent Variable: Qde_gordura_Y

15

Exemplo: Resultados do ajuste do modelo 3

Modelo 3: Regressão da quantidade de gordura (Y) sobre espessura da pele (X_1) e sobre o perímetro da coxa (X_2);

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	385,439	2	192,719	29,797	,000[a]
	Residual	109,951	17	6,468		
	Total	495,389	19			

a. Predictors: (Constant), Perim_coxa_X2, Esp_pele_X1

b. Dependent Variable: Qde_gordura_Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-19,174	8,361		-2,293	,035
	Esp_pele_X1	,222	,303	,219	,733	,474
	Perim_coxa_X2	,659	,291	,676	2,265	,037

a. Dependent Variable: Qde_gordura_Y

16

Exemplo: Resultados do ajuste do modelo 4

Modelo 4: Regressão da quantidade de gordura (Y) sobre espessura da pele (X_1), o perímetro da coxa (X_2) e perímetro braquial (X_3).

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	396,985	3	132,328	21,516	,000[a]
	Residual	98,405	16	6,150		
	Total	495,389	19			

a. Predictors: (Constant), Perim_braco_X3, Perim_coxa_X2, Esp_pele_X1

b. Dependent Variable: Qde_gordura_Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	117,085	99,782		1,173	,258
	Esp_pele_X1	4,334	3,016	4,264	1,437	,170
	Perim_coxa_X2	-2,857	2,582	-2,929	-1,106	,285
	Perim_braco_X3	-2,186	1,595	-1,561	-1,370	,190

a. Dependent Variable: Qde_gordura_Y

17

Modelo de Regressão Linear Múltipla (RLM):

(2) <u>Teste de comparabilidade de modelos (Teste F parcial)</u>

Definições e Notações:

Modelo 1:

- $SQReg(X_1)$ → soma de quadrados da regressão quando apenas X_1 está no modelo.
- $SQRes(X_1)$ → soma de quadrados dos resíduos quando apenas X_1 está no modelo.

Modelo 2:

- $SQReg(X_2)$ → soma de quadrados da regressão quando apenas X_2 está no modelo.
- $SQRes(X_2)$ → soma de quadrados dos resíduos quando apenas X_2 está no modelo.

18

Modelo de Regressão Linear Múltipla (RLM):

(2) <u>Teste de comparabilidade de modelos (Teste F parcial)</u>

Definições e notações:

Modelo 3:

- $SQReg(X_1,X_2)$ → soma de quadrados da regressão quando X_1 e X_2 estão incluídas no modelo.
- $SQRes(X_1,X_2)$ → soma de quadrados dos resíduos quando X_1 e X_2 estão incluídas no modelo.

Modelo 4:

- ✓ $SQReg(X_1,X_2,X_3)$ → soma de quadrados da regressão quando X_1,X_2 e X_3 estão incluídas no modelo.
- ✓ $SQRes(X_1,X_2,X_3)$ → soma de quadrados dos resíduos quando X_1,X_2 e X_3 estão incluías no modelo.

19

Observe, no exemplo, que a $SQRes(X_1,X_2)$=109,951 é menor do que aquela que contém apenas X_1 no modelo, $SQRes(X_1)$=143,120. A diferença é conhecida por *soma de quadrados extra de X_2 dado que X_1 já está incluído no modelo*, sendo denotada por $SQReg(X_2|X_1)$:

$SQReg(X_2/X_1)=SQRes(X_1) – SQRes(X_1,X_2)$

$SQReg(X_2/X_1)$=143,120 – 109,951= **33,169**

Esta redução na soma de quadrados dos resíduos (SQRes) é o resultado da adição de X_2 no modelo dado que X_1 já estava incluída no modelo. Esta soma de quadrados extra, denotada por $SQReg(X_2|X_1)$, mede a contribuição adicional quando X_2 é incluída no modelo dado que X_1 já estava incluída no modelo.

Alternativamente, podemos calcular a *soma de quadrados extra de X_2 dado X_1, isto é,* $SQReg(X_2/X_1)$, da seguinte forma:

$SQReg(X_2/X_1)=SQReg(X_1,X_2) – SQReg(X_1)$

$SQReg(X_2/X_1)$=385,439 – 352,270 = **33,169**

20

Analogamente, podemos calcular a *soma de quadrados extra de X_3 dado que X_1 e X_2* já estão incluídas no modelo, denotada por $SQReg(X_3/X_1,X_2)$, da seguinte forma:

$SQReg(X_3/X_1,X_2)=SQRes(X_1,X_2) – SQRes(X_1,X_2,X_3)$

$SQReg(X_3/X_1,X_2)$= 109,951 – 98,405 = **11,546**

Alternativamente, podemos calcular $SQReg(X_3/X_1,X_2)$ da seguinte forma:

$SQReg(X_3/X_1,X_2)=SQReg(X_1,X_2,X_3) – SQReg(X_1,X_2)$

$SQReg(X_3/X_1,X_2)$= 396,985 – 385,439 = **11,546**

21

Outra soma de quadrados extra de X_2 e X_3 dado que X_1 já está incluída no modelo, denotada por $SQReg(X_2,X_3/X_1)$:

$SQReg(X_2,X_3/X_1)=SQRes(X_1) – SQRes(X_1,X_2,X_3)$

$SQReg(X_2,X_3/X_1)$= 143,120 – 98,405 = **44,715**

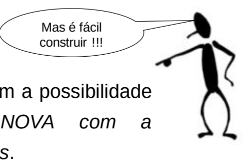
Alternativamente, podemos calcular $SQReg(X_2,X_3/X_1)$ da seguinte forma:

$SQReg(X_2,X_3/X_1)=SQReg(X_1,X_2,X_3) – SQReg(X_1)$

$SQReg(X_2,X_3/X_1)$= 396,985 – 352,270 = **44,715**

22

<u>Tabela ANOVA com a *Decomposição da SQReg em |Somas de Quadrados Extras (SQextras)*</u>:



➢ Alguns pacotes estatísticos oferecem a possibilidade de apresentação da *Tabela ANOVA com a decomposição da SQReg em SQextras*.

➢ Se as variáveis explicativas são incluídas <u>na ordem de seus índices</u>, as SQextras fornecidas na tabela são:

- $SQReg(X_1)$
- $SQReg(X_2/X_1)$
- $SQReg(X_3/X_1,X_2)$
- $SQReg(X_4/X_1,X_2,X_3)$

....

23

Tabela ANOVA com a decomposição da SQReg em somas de quadrados extras (SQextras):

➢ Se, por ex., queremos saber a contribuição $SQReg(X_3/X_1,X_2)$, devemos incluir as variáveis explicativas na ordem X_2, X_1, X_3:

- $SQReg(X_2)$
- $SQReg(X_1/X_2)$
- $SQReg(X_3/X_1,X_2)$

➢ Se, por ex., queremos saber a contribuição $SQReg(X_1/X_2,X_3)$, devemos incluir as variáveis explicativas na ordem X_2, X_3, X_1:

- $SQReg(X_2)$
- $SQReg(X_3/X_2)$
- $SQReg(X_1/X_2,X_3)$

24

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

Modelo de Regressão Linear Múltipla (RLM) Normal:

II) Teste de comparabilidade de modelos encaixados:
(Teste F parcial: Adição de 1 variável explicativa)

Exemplo (CASO 2): Com os dados de gordura corporal, vamos verificar se a adição da variável "perímetro braquial (X_3)" melhora significativamente a predição de Y, mantendo as demais variáveis no modelo.

Podemos excluir a variável "perímetro braquial (X_3)" do modelo ?

☐ Sim

☐ Não

25

Modelo de Regressão Linear Múltipla (RLM) Normal:

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

O modelo reduzido é tão adequado quanto o completo

❑ **Hipóteses a serem testadas:** $\begin{cases} H_0: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \\ H_1: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \end{cases}$

➢ Modelo reduzido de RLM de 2 vars explicativas (modelo 3):

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

$SQRes_0 = SQRes(X_1, X_2)$ e $q = 3$

➢ Modelo completo de RLM de 3 vars explicativas (modelo 4):

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

$SQRes_1 = SQRes(X_1, X_2, X_3)$ e $p = 4$

26

Modelo de Regressão Linear Múltipla (RLM) Normal:

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

❑ **Estatística de Teste:**

➢ Desse modo:

Modelo reduzido:

$SQRes_0 = SQRes(X_1, X_2) = 109{,}951$ e $q = 3$

Modelo completo:

$SQRes_1 = SQRes(X_1, X_2, X_3) = 98{,}405$ e $p = 4$

27

Modelo de Regressão Linear Múltipla (RLM) Normal:

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

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

$$F = \frac{(SQRes_0 - SQRes_1)/(p-q)}{SQRes_1/(n-p)} = \frac{(SQRes(X_1, X_2) - SQRes(X_1, X_2, X_3))/(p-q)}{SQRes(X_1, X_2, X_3)/(n-p)} \sim F_{p-q,\,n-q}$$

QMReg extra pela adição de X_3 dadas as vars X_1 e X_2.

$$F = \frac{QMReg(X_3/X_1, X_2)}{QMRes(X_1, X_2, X_3)}$$

QMRes para o modelo completo

28

Modelo de Regressão Linear Múltipla (RLM) Normal:

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

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

Valor observado de F:

$$f_{obs} = \frac{(SQRes(X_1, X_2) - SQRes(X_1, X_2, X_3))/(p-q)}{SQRes(X_1, X_2, X_3)/(n-p)} = \frac{(109{,}951 - 98{,}405)/(4-3)}{98{,}405/(20-4)} =$$

$$= \frac{11{,}546/1}{98{,}405/16} \cong 1{,}877$$

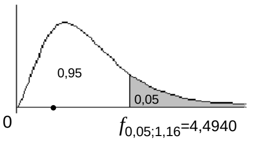
29

Modelo de Regressão Linear Múltipla (RLM):

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

❑ **Região crítica**:

$RC = \{ f \in \Re \;/\; f \geq 4{,}4940 \}$



❑ **Tomada de decisão:**

Como f_{obs}=1,877 $\notin$RC não há evidências para rejeitar H_0 ao nível de significância de 5%, ou seja, a variável "perímetro braquial (x_3)" pode ser excluída do modelo (β_3=0) que já contem as variáveis "espessura da pele(X_1)" e "perímetro da coxa(X_2)".

30

Tabela ANOVA com a decomposição da SQReg em somas de quadrados extras:

Tabela 1: Para os dados da gordura corporal, no caso de um modelo com 3 variáveis explicativas (X_1,X_2 e X_3):

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio
Regressão	**$SQReg(X_1,X_2,X_3)$=396,985**	**3**	**$QMReg(X_1,X_2,X_3)$=132,328**
X_1	$SQReg(X_1)$=352,270	1	$QMReg(X_1)$=352,270
X_2/X_1	$SQReg(X_2/X_1)$=33,169	1	$QMReg(X_2/X_1)$=33,169
$X_3/X_1,X_2$	$SQReg(X_3/X_1,X_2)$=11,546	1	$QMReg(X_3/X_1,X_2)$=11,546
Resíduos	**$SQRes(X_1,X_2,X_3)$=98,405**	**n-4=16**	**$QMRes(X_1,X_2,X_3)$=6,150**
Total	**$SQT(X_1,X_2,X_3)$=495,390**	**n-1=19**	

$$f_{obs} = \frac{QMReg(X_3/X_1,X_2)}{QMRes(X_1,X_2,X_3)} = \frac{11,546/1}{98,405/16} = \frac{11,546}{6,150} \cong 1,877$$

31

Modelo de Regressão Linear Múltipla (RLM) Normal:

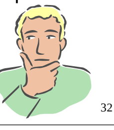
III) <u>Teste de comparabilidade de modelos (Teste F parcial: adição de um conjunto de variáveis)</u>

Exemplo (CASO 3): <u>Com os dados de gordura corporal</u>, vamos verificar se a adição simultânea do perímetro da coxa (X_2) e perímetro braquial (X_3) contribui significativamente para a predição de Y, mantendo a variável espessura da pele (X_1) no modelo.

Neste caso, podemos excluir as variáveis perímetros da coxa (X_2) e do braço (X_3)" do modelo ?

☐ Sim

☐ Não



32

Modelo de Regressão Linear Múltipla (RLM) Normal:

III) <u>Teste de comparabilidade de modelos (Teste F parcial: adição de um conjunto de variáveis)</u>

❑ **Hipóteses a serem testadas:**

O modelo reduzido é tão adequado quanto o completo

$$\begin{cases} H_0: \ Y = \beta_0 + \beta_1 X_1 + \varepsilon \\ H_1: \ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \end{cases} \qquad \begin{cases} H_0: \ \beta_2 = \beta_3 = 0 \\ H_1: \ \beta_j \neq 0 \text{ ao menos um j, j} = 2,3 \end{cases}$$

➢ Modelo reduzido de RLS de 1 var explicativa:

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$

$SQRes_0 = SQRes(X_1)$ e $q = 2$

➢ Modelo completo de RLM de 3 vars explicativas:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

$SQRes_1 = SQRes(X_1,X_2,X_3)$ e $p = 4$

33

Modelo de Regressão Linear Múltipla (RLM) Normal:

III) <u>Teste de comparabilidade de modelos (Teste F parcial: adição de um conjunto de variáveis)</u>

❑ **Estatística de Teste:**

➢ Desse modo:

<u>Modelo reduzido:</u>

$SQRes_0 = SQRes(X_1)$ =143,120 e q=2

<u>Modelo completo:</u>

$SQRes_1 = SQRes(X_1,X_2,X_3)$= 98,405 e p=4

34

Modelo de Regressão Linear Múltipla (RLM) Normal:

III) <u>Teste de comparabilidade de modelos (Teste F parcial: adição de um conjunto de variáveis)</u>

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

$$F = \frac{(SQRes_0 - SQRes_1)/(p-q)}{SQRes_1/(n-p)} = \frac{(SQRes(X_1) - SQRes(X_1,X_2,X_3))/(p-q)}{SQRes(X_1,X_2,X_3)/(n-p)} \sim F_{p-q,n-q}$$

QMReg extra pela adição de X_2 e X_3 dada a var. X_1.

$$F = \frac{QMReg(X_2,X_3/X_1)}{QMRes(X_1,X_2,X_3)}$$

QMRes para o modelo completo

35

Modelo de Regressão Linear Múltipla (RLM) Normal:

III) <u>Teste de comparabilidade de modelos (Teste F parcial: adição de um conjunto de variáveis)</u>

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

<u>Valor observado de F</u>:

$$f_{obs} = \frac{(SQRes(X_1) - SQRes(X_1,X_2,X_3))/(p-q)}{SQRes(X_1,X_2,X_3)/(n-p)} = \frac{(143,120 - 98,405)/(4-2)}{98,405/(20-4)} =$$

$$= \frac{44,715/2}{98,405/16} \cong 3,635$$

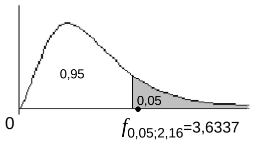
36

Modelo de Regressão Linear Múltipla (RLM) Normal:

III) <u>Teste de comparabilidade de modelos (Teste F parcial: adição de um conj. de vars)</u>

- **Região crítica**:

$$RC = \{ f \in \Re \ / \ f \geq 3{,}6337 \}$$



- **Tomada de decisão:**

Como f_{obs}=3,635 $\in$RC há evidências para rejeitar H_0 ao nível de significância de 5%, ou seja, as variáveis "perímetro da coxa (X_2)" e "perímetro braquial (X_3)" não podem ser excluídas.

37

Tabela ANOVA com a decomposição da SQReg em somas de quadrados extras:

Tabela 2: Para o mesmo modelo de 3 variáveis explicativas, segue a tabela ANOVA, onde na primeira etapa o modelo contemplava apenas 1 variável explicativa (X_1), e na segunda etapa, três variáveis explicativas (com a inclusão de X_2 e X_3).

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio
Regressão	**SQReg(X_1,X_2,X_3)=396,985**	**3**	**QMReg(X_1,X_2,X_3)=132,328**
X_1	SQReg(X_1)=352,270	1	QMReg(X_1)=352,270
X_2,X_3/X_1	SQReg(X_2,X_3/X_1)=44,715	2	QMReg(X_2,X_3/X_1)=22,358
Resíduos	**SQRes(X_1,X_2,X_3)=98,405**	**n-4=16**	**QMRes(X_1,X_2,X_3)=6,150**
Total	**SQT(X_1,X_2,X_3)=495,390**	**n-1=19**	

$$f_{obs} = \frac{QMReg(X_2, X_3/X_1)}{QMRes(X_1, X_2, X_3)} = \frac{44{,}715/2}{98{,}405/16} = \frac{22{,}358}{6{,}150} \cong 3{,}635$$

38

- **Exercício complementar:** <u>**Utilizando ainda os dados das medidas antropométricas.**</u>

Fazendo outras comparações de modelos, através do Teste F (1-*Hipóteses a serem testadas;* 2-*Estatística de teste;* 3-*Região Crítica;* 4-*Tomada de decisão*), qual modelo você escolheria ? Mostre todas as etapas de teste até a escolha do modelo final, além dos procedimentos/cálculos realizados. Escreva a equação do modelo final, descrevendo seus termos/variável(eis), no contexto do problema. Avalie para o modelo selecionado, a hipótese de normalidade dos erros (usando o "*QQ-Plot*" e o "*teste de Kolmogorov-Sminorv*") e a presença de outliers. Para tanto, utilize os resíduos estudentizados do modelo.

Resp.: Modelo c/ a var. "perímetro da coxa (X_2)"

39

Aula prática – Exercício 1 ("*Saídas*"): *Índice de distúrbio mental*

Um estudo no condado de Alachua, Flórida, investigou o relacionamento entre certos índices de saúde mental e diversas variáveis explicativas, tais como o escore dos eventos vividos (X_1) e posição socioeconômica (X_2). O interesse principal do estudo estava focado no índice de distúrbio mental (Y) que incorporou dimensões de sintomas psiquiátricos, incluindo aspectos de ansiedade e depressão. Escores maiores altos desse índice indicavam maior distúrbio mental. Com relação às duas variáveis explicativas mencionadas, os escores dos eventos vividos é uma medida composta da severidade dos principais eventos vividos que o indivíduo experimentou nos últimos três anos.

40

Aula prática – Exercício 1 ("*Saídas*"): *Índice de distúrbio mental* (*continuação*)

Esses eventos variavam de transtornos pessoais graves, como uma morte na família para eventos menos graves, como mudar-se de local de moradia. Assim essa medida, variou de 3 a 97 na amostra, sendo que um escore alto é indicativo de uma maior gravidade nos eventos vividos. Quanto a variável "posição sócio-econômica", é um índice composto baseado na ocupação, renda e nível educacional do indivíduo, mensurado numa escala que varia de 0 a 100, sendo que quanto maior o escore, maior o nível socioeconômico do indivíduo. Os dados do referido estudo são fornecidos na tabela a seguir:

41

Id	Distúrb. mental (Y)	Eventos vividos (X_1)	PSE (X_2)	Id	Distúrb. mental (Y)	Eventos vividos (X_1)	PSE (X_2)
1	17	46	84	**21**	27	60	70
2	19	39	97	**22**	28	97	89
3	20	27	24	**23**	28	37	50
4	20	3	85	**24**	28	30	90
5	20	10	15	**25**	28	13	56
6	21	44	55	**26**	28	40	56
7	21	37	78	**27**	29	5	40
8	22	35	91	**28**	30	59	72
9	22	78	60	**29**	30	44	53
10	23	32	74	**30**	31	35	38
11	24	33	67	**31**	31	95	29
12	24	18	39	**32**	31	63	53
13	25	81	87	**33**	31	42	7
14	26	22	95	**34**	32	38	32
15	26	50	40	**35**	33	45	55
16	26	48	52	**36**	34	70	58
17	26	45	61	**37**	34	57	16
18	27	21	45	**38**	34	40	29
19	27	55	88	**39**	41	49	3
20	27	45	56	**40**	41	89	75

42

Aula prática - Exercício 1 (*"Saídas"*): *Índice de distúrbio mental* (*continuação*)

a) Utilize o teste de comparabilidade de modelos (Teste F parcial), para responder a seguinte pergunta: Pelo menos uma das variáveis explicativas tem efeito estatisticamente significante ao nível de 5% ? ***OBS:*** *É necessário especificar a(s): 1º) Hipóteses a serem testadas; 2º) Estatística de teste; 3º)Região Crítica; 4º)Tomada de decisão*.

b) Ainda utilizando o teste F de comparabilidade de modelos (Teste F parcial) e um nível de significância de 5%, efetue as seguintes comparações:

1) $E(Y)=\beta_0+\beta_1X_1+\beta_2X_2$ *versus* $E(Y)=\beta_0+\beta_1X_1$

2) $E(Y)=\beta_0+\beta_1X_1+\beta_2X_2$ *versus* $E(Y)=\beta_0+\beta_2X_2$

43

Aula prática - Exercício 1 (*"Saídas"*): *Índice de distúrbio mental* (*continuação*)

c) Utilize outras medidas de qualidade do ajuste para fundamentar a escolha do modelo final. Escreva a equação do modelo final, descrevendo seus termos/variável (eis), no contexto do problema. Interprete as estimativas dos parâmetros do modelo que você selecionou, e avalie também a significância das relações encontradas usando o teste T, ao nível de 5%. Em sua opinião, o sentido das relações encontradas é o esperado ?

Resp.:

a) $f_{obs}=9{,}495$; b) Teste1: $f_{obs}=11{,}232$; Teste2: $f_{obs}=10{,}095$

<u>*As saídas do programa SPSS se encontram a seguir:*</u>

44

Aula prática – Exercício 1: Resultados do ajuste do modelo 1

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,582[a]	,339	,303	4,55644

a. Predictors: (Constant), pse_X2, ev_vividos_X1

b. Dependent Variable: dist_mental_Y

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	394,238	2	197,119	9,495	,000[a]
	Residual	768,162	37	20,761		
	Total	1162,400	39			

a. Predictors: (Constant), pse_X2, ev_vividos_X1

b. Dependent Variable: dist_mental_Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	28,230	2,174		12,984	,000
	ev_vividos_X1	,103	,032	,428	3,177	,003
	pse_X2	-,097	,029	-,451	-3,351	,002

a. Dependent Variable: dist_mental_Y

45

Aula prática – Exercício 1: Resultados do ajuste do modelo 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,372[a]	,139	,116	5,13336

a. Predictors: (Constant), ev_vividos_X1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	161,048	1	161,048	6,112	,018[a]
	Residual	1001,352	38	26,351		
	Total	1162,400	39			

a. Predictors: (Constant), ev_vividos_X1

b. Dependent Variable: dist_mental_Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	23,309	1,807		12,901	,000
	ev_vividos_X1	,090	,036	,372	2,472	,018

a. Dependent Variable: dist_mental_Y

46

Aula prática – Exercício 1: Resultados do ajuste do modelo 3

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,399[a]	,159	,137	5,07249

a. Predictors: (Constant), pse_X2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	184,654	1	184,654	7,177	,011[a]
	Residual	977,746	38	25,730		
	Total	1162,400	39			

a. Predictors: (Constant), pse_X2

b. Dependent Variable: dist_mental_Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	32,172	1,988		16,186	,000
	pse_X2	-,086	,032	-,399	-2,679	,011

a. Dependent Variable: dist_mental_Y

47

Aula prática – Exercício 1: Resultados do ajuste do modelo 4 (modelo nulo)

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	29811,600	1	29811,600	1000,217	,000[a]
	Residual	1162,400	39	29,805		
	Total	30974,000[b]	40			

a. Predictors: Intercept

b. This total sum of squares is not corrected for the constant because the constant is zero for regression through the origin.

c. Dependent Variable: dist_mental_Y

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Intercept	27,300	,863	,981	31,626	,000

a. Dependent Variable: dist_mental_Y

b. Linear Regression through the Origin

48

Aula prática – Exercício 2: <u>Voltando ao exemplo de aplicação</u>: *Modelo de Regressão Linear Múltipla* (*RLM*)

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)

49

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				

50

Aula prática - Exercício 2 (*continuação*):

Considerando os dados da tabela 1, 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, Região Crítica e Tomada de decisão)*.

51

Aula prática - Exercício 2 (*continuação*):

d) Avalie se a "idade" e o "tempo de serviço" influenciam no salário dos funcionários da empresa. 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 compare graficamente (e por meio de alguma medida apropriada) os salários observados e os estimados.

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

52

Observações:

- ➢ Essa forma de análise (*Teste de comparação de modelos* ou *Teste F parcial*) é uma das melhores ferramentas para seleção de modelos de regressão.
- ➢ Um outro método menos rigoroso para comparar e selecionar modelos é através do coeficiente de múltipla determinação ajustado ($R^2_{ajustado}$).

53

Avisos:

- ✓ Fazer a **3ª Lista de Exercícios** (*Regressão Linear Múltipla*) da disciplina "Modelos Lineares I", proposta pelo Prof. Dr. José Rodrigo de Moraes.
- ✓ Recomenda-se a leitura do livro-texto (ver ementa da disciplina);
- ✓ Para as questões que requerem auxílio computacional deve ser utilizado o *RStudio*.

54