Modelos Lineares I

Regressão Linear Múltipla (RLM):

Observações discrepantes e influentes

(30ª e 31ª Aulas)

Professor: Dr. José Rodrigo de Moraes

Universidade Federal Fluminense (UFF)

Departamento de Estatística (GET)

1

Modelo de Regressão Linear Múltipla Normal:

Introdução – Influência no ajuste do modelo:

Em situações práticas, um pequeno subconjunto de observações pode exercer influência desproporcional no ajuste do modelo de regressão.

- Será preocupante se pequenas pertubações nestas observações produzirem mudanças substanciais nas estimativas dos parâmetros do modelo (dependência do modelo).

Procedimento: Localizar essas observações e avaliar o impacto no modelo:

- Se essas observações influentes forem valores "ruins", elas podem ser eliminadas;
- E se controlarem propriedades importantes do modelo, é bom saber disso !!!

2

Análise Gráfica dos Resíduos do Modelo de RLS:

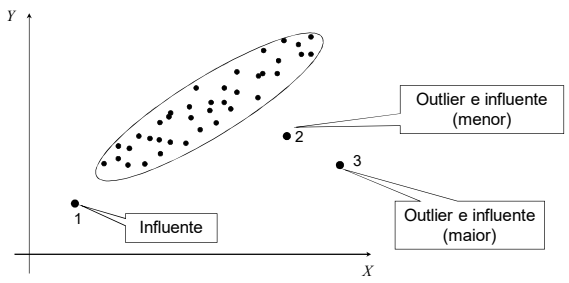
***Outliers (valores discrepantes ou atípicos) -Recordando*:**

- Outliers são observações discrepantes que apresentam resíduos que são consideravelmente superiores aos resíduos de outras observações.
- Os efeitos do *outlier* podem ser moderados ou extremos (dependendo do local onde se encontra), em termos da magnitude da sua influência sobre a estimação dos parâmetros. *Vejamos o exemplo a seguir*:

Observações discrepantes (outliers) *versus* observações influentes. **OBS:** Uma observação pode ser simultamente um outlier e influente !!!

3

Ilustração: Modelo de RLS com três observações extraordinárias / extremas (influentes e Outliers)



4

Comentários - Ilustração:

❑Embora o Ponto 1 seja extremo, não vai influenciar indevidamente a inclinação e o intercepto do modelo.

❑Já os Pontos 2 e 3 são pontos influentes, mas o Ponto 3 é um ponto altamente influente na inclinação e intercepto do modelo. Embora o Ponto 2 seja também influente, sua influência é menor do que a do Ponto 3.

❑ Os pontos influentes são susceptíveis de serem encontrados em "áreas" onde pouco ou nenhum outro dado foi coletado. Esses pontos podem ser muito bem ajustados, mas em *detrimento* do ajuste de outros dados.

5

Comentários (*continuação*) - Ilustração:

❑Em termos de ajuste do modelo, os pontos (observações) com alta alavancagem podem ser "*bons*" ou "*ruins*": O Ponto 1 pode reduzir as estimativas das variâncias dos estimadores de β_0 e β_1, enquanto o Ponto 3 pode alterar drasticamente o ajuste do modelo.

❑Se o Ponto 3 não é resultado de erro de informação ou classificação, então o pesquisador deve escolher qual modelo adotar. Normalmente, o modelo que melhor se ajusta a maioria dos dados deve ser o preferido até que observações adicionais possam ser observadas em outras áreas.

6

Medidas de diagnóstico de influência:

A seguir são apresentadas algumas das principais medidas usadas para a identificação de observações ditas influentes, entre elas destaca-se a "*Distância de Cook*".

Matriz de projeção (H):

Muito usada nas técnicas de diagnostico de análise de regressão linear: os elementos $h_{11}, h_{22}, ..., h_{nn}$ da diagonal da matriz **H=X(X`X)$^{-1}$X`** são denominadas medidas de alavancagem ("*leverage*").

- A inspeção de tais elementos podem revelar pontos (observações) que são potencialmente influentes na determinação dos coeficientes do modelo (ajuste do modelo), em virtude da sua localização no espaço X (*ilustração*).

7

Observações: Matriz de Projeção (ou matriz Hat):

- Os valores h_{ii} *(medidas de alavancagem="leverage")* da matriz H, tal que $\frac{1}{n} \le h_{ii} \le 1$, refletem parcialmente a influência de cada observação no ajuste do modelo.
- $h_{ii} \rightarrow$ Peso da observação Y_i na determinação do modelo ajustado.
- Em geral:
  - $h_{ii} \cong 1/n \rightarrow$ a obs. i tem baixo potencial para influenciar o ajuste do modelo.
  - $h_{ii} \cong 1 \rightarrow$ a obs. i tem alto potencial para influenciar o ajuste do modelo.

8

Observações: Matriz de Projeção (ou matriz Hat):

- Como a $\sum_{i=1}^{n} h_{ii} = p$, supondo que todas as observações exerçam a mesma influência sobre os valores ajustados ($\hat{\mathbf{Y}} = \mathbf{HY}$), espera-se que h_{ii} esteja próximo de p/n.
- **Valor crítico:** $h_{ii} > 2p/n$ *(p é o número de parâmetros)* $\rightarrow$ *a observação i é uma observação (ponto) de alta alavancagem.*
- O SPSS calcula *medidas centradas de alavancagem* (h_{ii}*), que variam no intervalo $0 \le h_{ii}^* \le 1 - \frac{1}{n}$, dadas por:

$$h_{ii}^* = h_{ii} - \frac{1}{n}$$

$$h_{ii} = h_{ii}^* + \frac{1}{n}, \quad \text{onde: } n \rightarrow \text{tamanho da amostra}$$

$h_{ii}^* < 0{,}20$	$\rightarrow$ baixa alavancagem
$0{,}20 \le h_{ii}^* \le 0{,}50$	$\rightarrow$ moderada alavancagem
$h_{ii}^* > 0{,}50$	$\rightarrow$ alta alavancagem

9

Resíduos estudentizados (r^S) – *Detecção de outliers:*

- Vimos que os resíduos estudentizados são considerados medidas mais adequadas para o exame de valores discrepantes (*outliers*), pois a variância dos resíduos e_i`s não é constante. Com efeito, e_i tem distribuição normal de média 0 e variância VAR(e_i)= σ^2 (1 – h_{ii}). Os resíduos estudentizados, denotados por r^S, são definidos abaixo:

$$r_i^S = \frac{e_i}{\sqrt{\hat{\sigma}^2 \cdot (1 - h_{ii})}}; \quad \forall\, i = 1, 2, ..., n$$

onde: h_{ii} é o i-ésimo elemento da diagonal da matriz de projeção H=X(X`X)$^{-1}$X`.

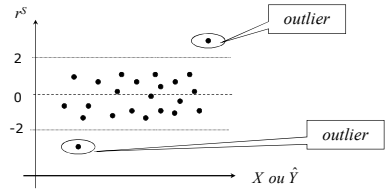
10

Análise Gráfica dos Resíduos do Modelo de RLS:

Resíduos estudentizados

Critério para identificação de outliers:

$|r_i^S| > 2 \rightarrow$ a unidade i é um ***outlier***.



OBS: 95% dos resíduos devem estar no intervalo [-2,+ 2].

11

Modelo de Regressão Linear Múltipla Normal:

Outra medida de influência:

- A informação de alavancagem (h_{ii}) reflete parcialmente a influência de uma observação (LEE, 1987).
- Para verificar a completa influência da i-ésima observação, torna-se necessário comparar as estimativas $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\beta}}_{(i)}$, esta última obtida quando a observação i é excluída da análise estatística (*avaliação do impacto*).

12

Modelo de Regressão Linear Múltipla Normal:

Distância de Cook:

Medida global de diagnóstico da influência de uma observação i sobre as estimativas dos parâmetros, que combina os resíduos estudentizados e as medidas de alavancagem h_{ii}.

- Foi proposta por Cook (1977), como uma medida da distância ao quadrado entre as estimativas dos parâmetros baseadas em todas n observações ($\hat{\boldsymbol{\beta}}$) e as estimativas obtidas com a exclusão da i-ésima observação ($\hat{\boldsymbol{\beta}}_{(i)}$). A *Distância de Cook*, denotada por D_i, tem a seguinte expressão:

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)` (\boldsymbol{X}`\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{p \cdot \hat{\sigma}^2}; \quad i = 1,2,...,n$$

13

Distância de Cook (*continuação*):

$\hat{\boldsymbol{\beta}}_{(i)} = \left(\boldsymbol{X}_{(i)}`\boldsymbol{X}_{(i)}\right)^{-1} \boldsymbol{X}_{(i)}`\boldsymbol{Y}_{(i)} \rightarrow$ Estimativa de $\boldsymbol{\beta}$ obtida pela exclusão da observação i.

❑ A *Distância de Cook* também pode ser escrita por:

$$D_i = \frac{\left(\hat{\boldsymbol{Y}}_{(i)} - \hat{\boldsymbol{Y}}\right)` \left(\hat{\boldsymbol{Y}}_{(i)} - \hat{\boldsymbol{Y}}\right)}{p \cdot \hat{\sigma}^2}; \quad i = 1,2,...,n$$

Pergunta: O que concluir sobre D_i, a partir das expressões acima?

14

Observações: Distância de Cook:

❑ Pode-se demonstrar ainda que a estatística D_i (Cordeiro & Neto, 2004), pode ser calculada por:

$$D_i = \frac{\left(r_i^S\right)^2}{p} \frac{h_{ii}}{1 - h_{ii}}; \quad i = 1,2,...,n$$

- Medida global de quão atípica a observação i se apresenta no ajuste do modelo.
- Cada componente (ou ambos) pode contribuir para um grande valor da estatística D_i.

15

Distância de Cook (*continuação*):

Observações:

❑ Valores grande da estatística D_i indicam observações que influenciam bastante as estimativas dos parâmetros e as predições.

❑ **Regra prática** (valor crítico=1):

- $D_i > 1 \rightarrow$ a obs. i é excessivamente influente.
- *Em geral, se o maior valor de D_i for muito inferior a 1, então a eliminação de qualquer observação, não vai alterar substancialmente as estimativas dos parâmetros (Cordeiro & Neto, 2004).*
- *Entretanto, nós podemos examinar quaisquer observações cujo valor de D_i seja extraordinariamente maior em relação aos demais valores de D_i.*

16

Observações discrepantes ou influentes:

Métodos Gráficos:

Para detectar problemas com o ajuste do modelo relacionada à *presença de observações discrepantes e influentes*, recomenda-se a utilização dos gráficos de dispersão entre:

- ✓ r_i^S *e a ordem das observações;*
- ✓ h_{ii} *e a ordem das observações;*
- ✓ D_i *e a ordem das observações;*
- ✓ D_i *e* r_i^S.

17

Exemplo 1: Modelo de Regressão Linear Simples

Os dados apresentados na tabela 1 a seguir se referem a um estudo sobre o desempenho de n=22 alunos, de uma escola técnica, matriculados na disciplina de matemática e estatística.

O objetivo do estudo é avaliar se alunos com melhores desempenhos em Matemática tendem apresentar melhores desempenhos na disciplina Estatística.

As notas para ambas as disciplinas variam de 0 a 10.

18

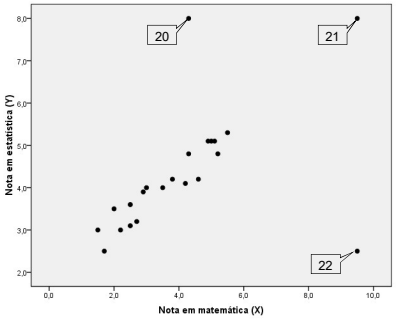
Tabela 1: Nota de matemática (X) versus Nota de Estatística (Y)

Aluno	Nota de matemática	Nota de estatística
1	1,5	3,0
2	1,7	2,5
3	2,0	3,5
4	2,2	3,0
5	2,5	3,1
6	2,5	3,6
7	2,7	3,2
8	2,9	3,9
9	3,0	4,0
10	3,5	4,0
11	3,8	4,2
12	4,2	4,1
13	4,3	4,8
14	4,6	4,2
15	4,9	5,1
16	5,0	5,1
17	5,1	5,1
18	5,2	4,8
19	5,5	5,3
20	4,3	8,0
21	9,5	8,0
22	9,5	2,5

19

Exemplo: Análise gráfica (n=22)

Figura 1: Gráfico de Dispersão entre as notas de matemática (X) e estatística (Y)



20

Exemplo (*1º Ajuste*): Resultados do ajuste do modelo n=19+obs(20)+obs(21)+obs(22)= 22 observações

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,499[a]	,249	,212	1,2972

a. Predictors: (Constant), X_nota_mat

b. Dependent Variable: Y_nota_est

Coefficients[a]

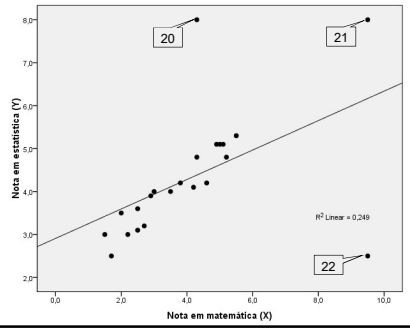
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2,909	,613		4,746	,000
	X_nota_mat	,343	,133	,499	2,577	,018

a. Dependent Variable: Y_nota_est

21

Exemplo: Análise gráfica (n=22)

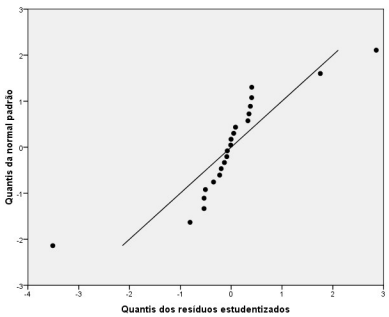
Figura 2: Gráfico de Dispersão entre as notas de matemática (X) e estatística (Y), incluindo a o modelo de regressão.



22

Exemplo: Análise gráfica (n=22)

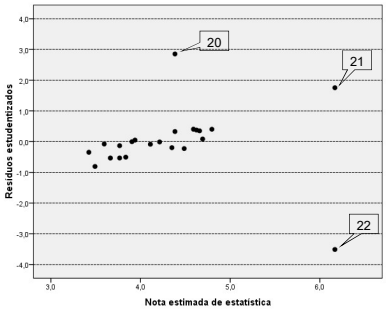
Figura 3: QQ-Plot (normalidade) para os resíduos estudentizados do modelo



23

Exemplo: Análise gráfica (n=22)

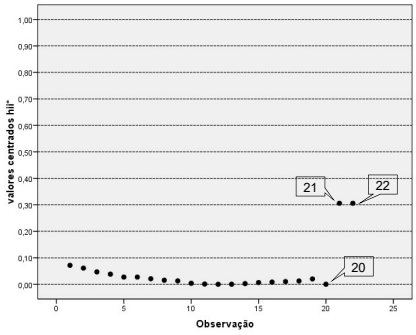
Figura 4: Gráfico de Dispersão entre a nota estimada de estatística e os resíduos estudentizados do modelo.



24

Exemplo: Análise gráfica (n=22)

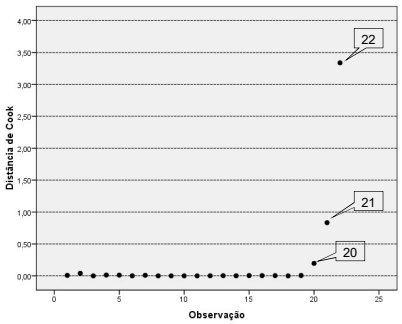
Figura 5: Gráfico de Dispersão entre a ordem das observações e as medidas centradas de alavancagem h_{ii}*



25

Exemplo: Análise gráfica (n=22)

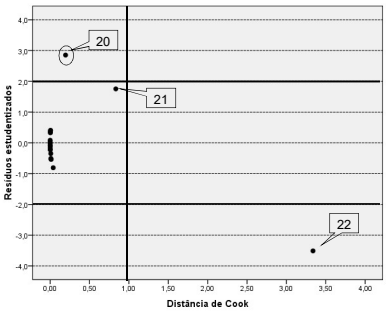
Figura 6: Gráfico de Dispersão entre a ordem das observações e as medidas da distância de Cook



26

Exemplo: Análise gráfica (n=22)

Figura 7: Gráfico de Dispersão entre as medidas da distância de Cook e os resíduos estudentizados.



27

Exemplo (*2º Ajuste*): Resultados do ajuste do modelo após a exclusão da "obs(20) = outlier".

n=19+ obs(21)+obs(22)= 21 observações

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,591[a]	,349	,314	1,0247

a. Predictors: (Constant), X_nota_mat

b. Dependent Variable: Y_nota_est

Coefficients[a]

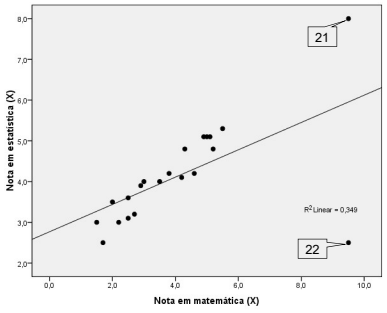
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2,768	,486		5,699	,000
	X_nota_mat	,335	,105	,591	3,190	,005

a. Dependent Variable: Y_nota_est

28

Exemplo: Análise gráfica (n=21)

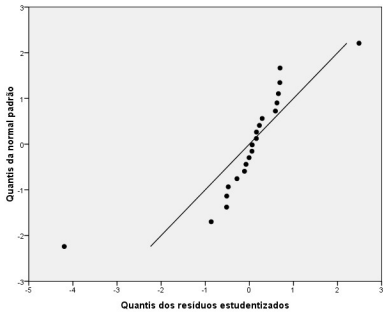
Figura 1: Gráfico de Dispersão entre as notas de matemática (X) e estatística (Y), incluindo a o modelo de regressão.



29

Exemplo: Análise gráfica (n=21)

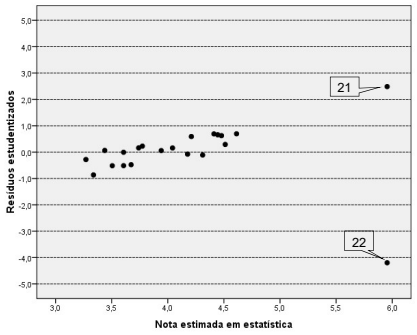
Figura 2: QQ-Plot (normalidade) para os resíduos estudentizados do modelo



30

Exemplo: Análise gráfica (n=21)

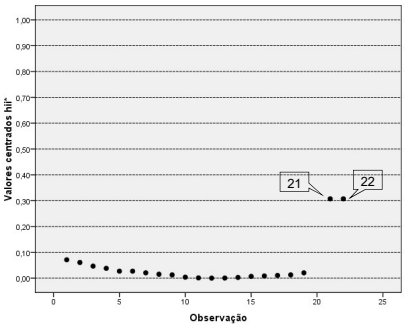
Figura 3: Gráfico de Dispersão entre a nota estimada de estatística e os resíduos estudentizados do modelo.



31

Exemplo: Análise gráfica (n=21)

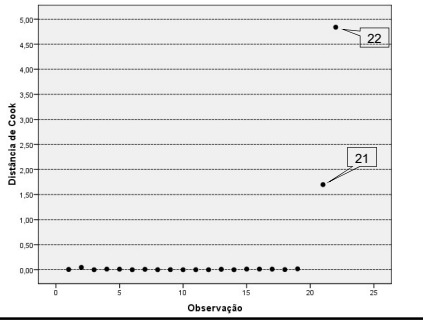
Figura 4: Gráfico de Dispersão entre a ordem das observações e as medidas centradas de alavancagem h_{ii}*



32

Exemplo: Análise gráfica (n=21)

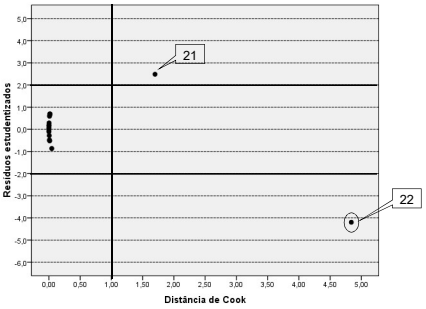
Figura 5: Gráfico de Dispersão entre a ordem das observações e as medidas da distância de Cook



33

Exemplo: Análise gráfica (n=21)

Figura 6: Gráfico de Dispersão entre as medidas da distância de Cook e os resíduos estudentizados.



34

Exemplo (*3º Ajuste*): Resultados do ajuste do modelo após a exclusão da "obs(22)=influente", além da "obs(20)=outlier". n=19+ obs(21)= 20 observações

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,973[a]	,947	,945	,2849

a. Predictors: (Constant), X_nota_mat

b. Dependent Variable: Y_nota_est

Coefficients[a]

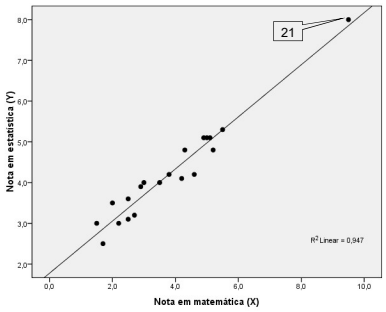
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1,775	,150		11,815	,000
	X_nota_mat	,640	,036	,973	18,015	,000

a. Dependent Variable: Y_nota_est

35

Exemplo: Análise gráfica (n=20)

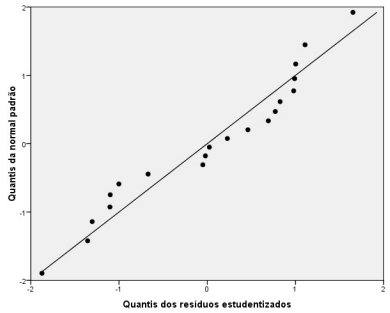
Figura 1: Gráfico de Dispersão entre as notas de matemática (X) e estatística (Y), incluindo a o modelo de regressão



36

Exemplo: Análise gráfica (n=20)

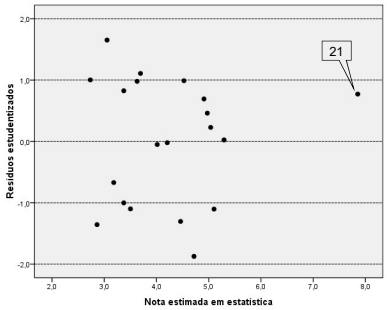
Figura 2: QQ-Plot (normalidade) para os resíduos estudentizados do modelo



37

Exemplo: Análise gráfica (n=20)

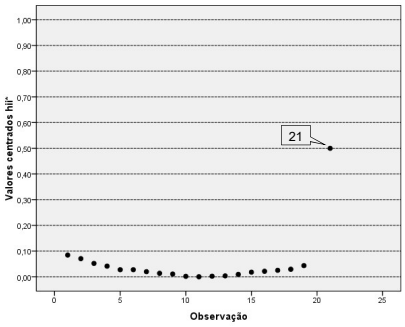
Figura 3: Gráfico de Dispersão entre a nota estimada de estatística e os resíduos estudentizados do modelo.



38

Exemplo: Análise gráfica (n=20)

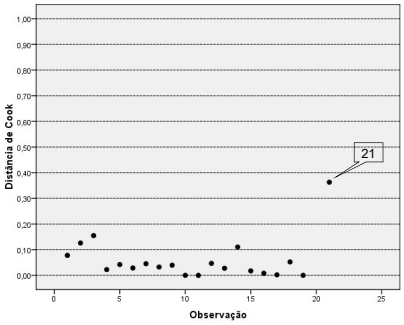
Figura 4: Gráfico de Dispersão entre a ordem das observações e as medidas centradas de alavancagem h_{ii}*



39

Exemplo: Análise gráfica (n=20)

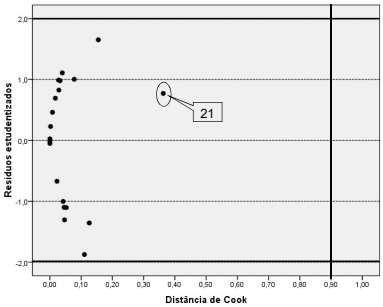
Figura 5: Gráfico de Dispersão entre a ordem das observações e as medidas da distância de Cook



40

Exemplo: Análise gráfica (n=20)

Figura 6: Gráfico de Dispersão entre as medidas da distância de Cook e os resíduos estudentizados.



41

Exemplo (*4º Ajuste*): Resultados do ajuste do modelo após a exclusão da "obs(21)", além de "obs(20)=outlier" e "obs(22)=influente".

n= 19 observações

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,943[a]	,890	,883	,2883

a. Predictors: (Constant), X_nota_mat

Coefficients[a]

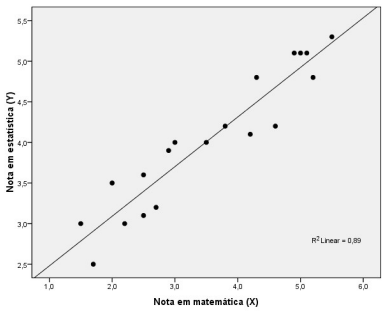
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,869	,196		9,543	,000
	X_nota_mat	,611	,052	,943	11,705	,000

a. Dependent Variable: Y_nota_est

42

Exemplo: Análise gráfica (n=19)

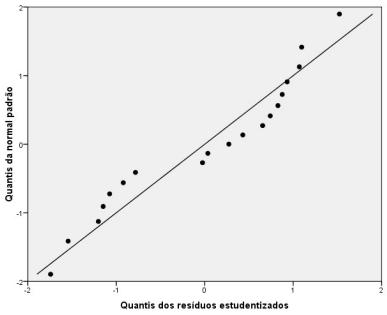
Figura 1: Gráfico de Dispersão entre as notas de matemática (X) e estatística (Y), incluindo a o modelo de regressão



43

Exemplo: Análise gráfica (n=19)

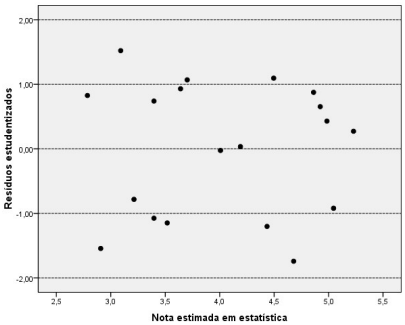
Figura 2: QQ-Plot (normalidade) para os resíduos estudentizados do modelo



44

Exemplo: Análise gráfica (n=19)

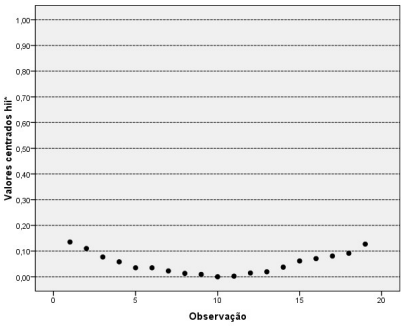
Figura 3: Gráfico de Dispersão entre a nota estimada de estatística e os resíduos estudentizados do modelo.



45

Exemplo: Análise gráfica (n=19)

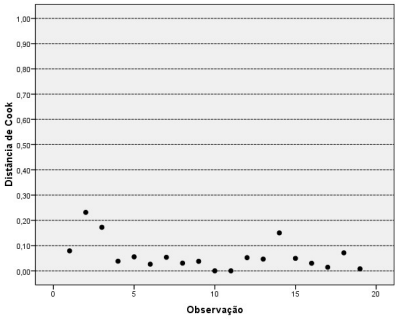
Figura 4: Gráfico de Dispersão entre a ordem das observações e as medidas centradas de alavancagem h_{ii}*



46

Exemplo: Análise gráfica (n=19)

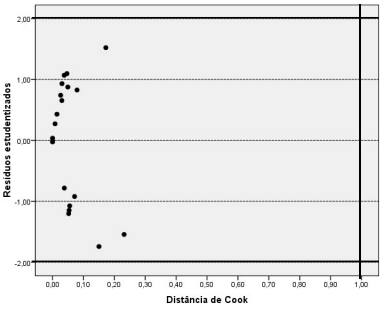
Figura 5: Gráfico de Dispersão entre a ordem das observações e as medidas da distância de Cook



47

Exemplo: Análise gráfica (n=19)

Figura 6: Gráfico de Dispersão entre as medidas da distância de Cook e os resíduos estudentizados.



48

Exemplo 1:

Resumo com os principais resultados dos 4 ajustes:

Ajustes	n	β_0	β_1	*P-valor*(β_1)	R^2	R^2*ajust*
1º) Todas as obs	22	2,909	0,343	0,018	24,9%	21,2%
2º) Excluindo a **obs(20)**	21	2,768	0,335	0,005	34,9%	31,4%
3º) Excluindo as obs(20) e **obs(22)**	20	1,775	0,640	<0,001	94,7%	94,5%
4º) Excluindo as obs(20), obs(22) e **obs(21)**	19	1,869	0,611	<0,001	89,0%	88,3%

49

Exemplo 2: Modelo de Regressão Linear Múltipla

Os dados apresentados na tabela a seguir se referem a um estudo sobre a duração do AME (Y), em meses, realizado com uma amostra de n=21 mães atendidas num determinado hospital.

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

- ✓ Anos de estudo da mãe (X_1)
- ✓ Tempo de orientação (em min) voltada ao manejo (X_2).
- ✓ Horas de trabalho (X_4)
- ✓ Tempo de casada (em anos) (X_5)

OBS: *A variável X_3 não foi considerada na análise !!!*

50

Banco de Dados Reduzido: ***Modelo RLM Normal (n=21 mães)*:**

Mãe	AME	Anos de estudo	Tempo de orientação (manejo)	Horas de trabalho	Tempo de casada
1	2,1	6	71	23	1
2	3,4	6	65	35	6
3	3,6	6	81	24	7
4	1,7	6	50	27	1
5	1,8	6	46	49	1
6	3,2	6	74	50	8
7	2,6	6	64	52	5
8	2,9	6	63	64	7
9	2,3	6	54	37	5
10	1,6	5	47	49	1
11	1,5	6	47	85	3
12	2,7	6	70	80	9
13	4,1	6	117	74	12
14	2,0	6	46	76	5
15	2,3	6	57	81	6
16	3,6	6	89	60	10
17	3,3	6	86	52	12
18	1,8	6	50	64	6
19	1,7	6	56	18	5
20	2,4	5	75	32	9
21	3,0	6	84	64	6

51

Exemplo 2: Resultados do ajuste do modelo 1

Modelo 1: Modelo completo (4 variáveis explicativas)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9,917	4	2,479	20,081	,000[a]
	Residual	1,975	16	,123		
	Total	11,892	20			

a. Predictors: (Constant), Tempo_casado, Anos_estudo, Horas_trabalho, Tempo_orientação_manejo

b. Dependent Variable: Duraçao_AME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-2,090	1,378		-1,517	,149		
	Anos_estudo	,486	,256	,212	1,902	,075	,832	1,202
	Tempo_orientação_manejo	,023	,007	,544	3,127	,007	,343	2,916
	Horas_trabalho	-,004	,004	-,112	-,965	,349	,766	1,306
	Tempo_casado	,083	,040	,361	2,059	,056	,338	2,959

a. Dependent Variable: Duraçao_AME

52

Exemplo 2: Resultados do ajuste do modelo 2

Modelo 2: Modelo com 3 variáveis explicativas

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9,802	3	3,267	26,573	,000[a]
	Residual	2,090	17	,123		
	Total	11,892	20			

a. Predictors: (Constant), Tempo_casado, Anos_estudo, Tempo_orientação_manejo

b. Dependent Variable: Duraçao_AME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-1,930	1,365		-1,414	,175		
	Anos_estudo	,410	,243	,179	1,690	,109	,920	1,087
	Tempo_orientação_manejo	,026	,007	,603	3,709	,002	,391	2,557
	Tempo_casado	,067	,036	,290	1,826	,085	,411	2,433

a. Dependent Variable: Duraçao_AME

53

Exemplo 2: Resultados do ajuste do modelo 3

Modelo 3: Modelo com 2 variáveis explicativas

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9,451	2	4,726	34,841	,000[a]
	Residual	2,441	18	,136		
	Total	11,892	20			

a. Predictors: (Constant), Tempo_casado, Tempo_orientação_manejo

b. Dependent Variable: Duraçao_AME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	,312	,336		,927	,366		
	Tempo_orientação_manejo	,028	,007	,666	4,006	,001	,413	2,423
	Tempo_casado	,063	,038	,272	1,635	,119	,413	2,423

a. Dependent Variable: Duraçao_AME

54

Exemplo 2: Resultados do ajuste do modelo 4

Modelo 4: Modelo com apenas 1 variável explicativa

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9,088	1	9,088	61,580	,000[a]
	Residual	2,804	19	,148		
	Total	11,892	20			

a. Predictors: (Constant), Tempo_orientação_manejo

b. Dependent Variable: Duraçao_AME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,099	,324		,307	,762		
	Tempo_orientação_manejo	,037	,005	,874	7,847	,000	1,000	1,000

a. Dependent Variable: Duraçao_AME

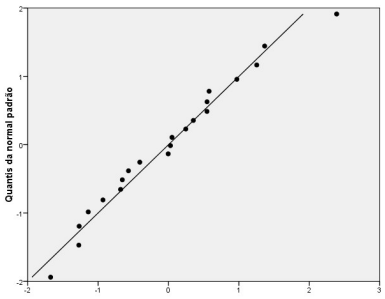
Modelo final : $\hat{Y}_i = 0{,}099 + 0{,}037\, X_i$; $i = 1{,}2{,}...{,}\ 21$

$R^2 = 76{,}4\%$

55

Exemplo 2: Análise dos resíduos para o modelo 4

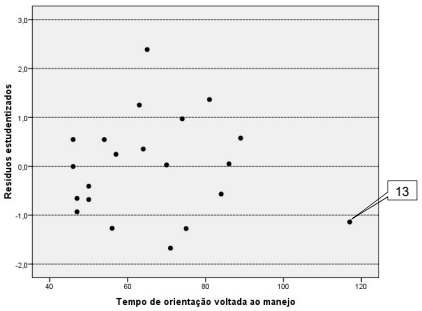
Figura 1: QQ-Plot (normalidade) para os resíduos estudentizados do modelo



56

Exemplo 2: Análise dos resíduos para o modelo 4

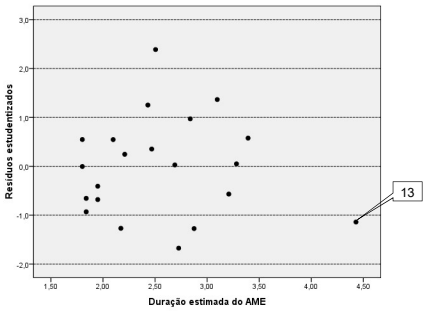
Figura 2: Gráfico de Dispersão entre o tempo de orientação voltada ao manejo e os resíduos estudentizados do modelo.



57

Exemplo 2: Análise dos resíduos para o modelo 4

Figura 3: Gráfico de Dispersão entre a duração estimada do AME e os resíduos estudentizados do modelo.



58

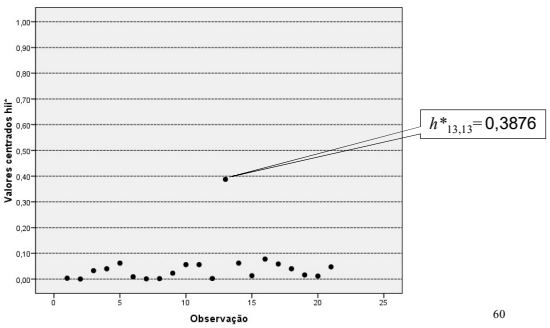
Medidas de Influência para o modelo 4.

Mãe i	Dist. de Cook (D_{ii})	$h_{ii}^*=h_{ii}-1/n$	h_{ii}
1	0,0753	0,0033	0,0510
2	0,1434	0,0002	0,0479
3	0,0813	0,0326	0,0802
4	0,0222	0,0400	0,0876
5	0,0000	0,0620	0,1096
6	0,0282	0,0090	0,0566
7	0,0032	0,0008	0,0484
8	0,0406	0,0016	0,0492
9	0,0113	0,0227	0,0704
10	0,0249	0,0560	0,1037
11	0,0501	0,0560	0,1037
12	0,0000	0,0021	0,0497
13	0,5006	0,3876	0,4352
14	0,0184	0,0620	0,1096
15	0,0019	0,0130	0,0606
16	0,0238	0,0777	0,1254
17	0,0001	0,0586	0,1062
18	0,0080	0,0400	0,0876
19	0,0546	0,0159	0,0636
20	0,0510	0,0114	0,0591
21	0,0170	0,0473	0,0949

59

Exemplo 2: Análise dos resíduos para o modelo 4

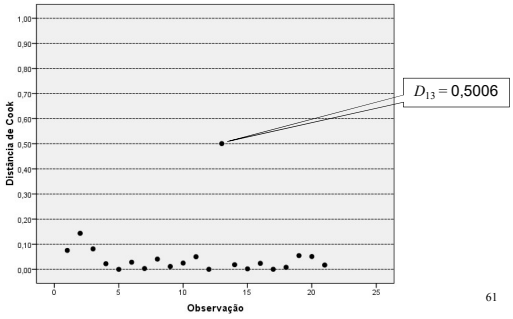
Figura 4: Gráfico de Dispersão entre a ordem das observações e as medidas centradas de alavancagem h_{ii}^*



60

Exemplo 2: Análise dos resíduos para o modelo 4

Figura 5: Gráfico de Dispersão entre a ordem das observações e as medidas da distância de Cook



61

Comentários / Perguntas:

Como nenhum dos valores de D_i excede a 1, não existe evidência de observações fortemente influentes no conjunto de dados.

Entretanto, para checar tal conclusão, a *obs* 13 (maior valor de D_i) será excluída, e o modelo reajustado (n=20 observação).

Perguntas:

1) *O mesmo modelo será selecionado?*
2) *Houve alteração nas estimativas dos parâmetros?*
3) *As hipóteses básicas do modelo permaneceram válidas?*

62

Exemplo: Resultados do ajuste do modelo 1

Modelo 1: Modelo completo (4 variáveis explicativas)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7,603	4	1,901	16,070	,000[a]
	Residual	1,774	15	,118		
	Total	9,378	19			

a. Predictors: (Constant), Tempo_casado, Anos_estudo, Horas_trabalho, Tempo_orientação_manejo

b. Dependent Variable: Duraçao_AME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-2,857	1,471		-1,942	,071		
	Anos_estudo	,546	,254	,248	2,146	,049	,943	1,061
	Tempo_orientação_manejo	,030	,009	,608	3,362	,004	,386	2,592
	Horas_trabalho	-,002	,004	-,070	-,530	,604	,734	1,363
	Tempo_casado	,069	,041	,307	1,685	,113	,379	2,638

a. Dependent Variable: Duraçao_AME

63

Exemplo: Resultados do ajuste do modelo 2

Modelo 2: Modelo com 3 variáveis explicativas

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7,570	3	2,523	22,336	,000[a]
	Residual	1,808	16	,113		
	Total	9,378	19			

a. Predictors: (Constant), Tempo_casado, Anos_estudo, Tempo_orientação_manejo

b. Dependent Variable: Duraçao_AME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-2,872	1,437		-1,998	,063		
	Anos_estudo	,514	,242	,234	2,128	,049	,997	1,003
	Tempo_orientação_manejo	,032	,008	,651	4,132	,001	,485	2,062
	Tempo_casado	,059	,035	,262	1,664	,116	,486	2,058

a. Dependent Variable: Duraçao_AME

64

Exemplo: Resultados do ajuste do modelo 3

Modelo 3: Modelo com 2 variáveis explicativas

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7,257	2	3,629	29,090	,000[a]
	Residual	2,120	17	,125		
	Total	9,378	19			

a. Predictors: (Constant), Tempo_orientação_manejo, Anos_estudo

b. Dependent Variable: Duraçao_AME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-3,066	1,505		-2,037	,057		
	Anos_estudo	,504	,254	,229	1,986	,063	,998	1,002
	Tempo_orientação_manejo	,041	,006	,839	7,270	,000	,998	1,002

a. Dependent Variable: Duraçao_AME

65

Exemplo: Resultados do ajuste do modelo 4

Modelo 4: Modelo com apenas 1 variável explicativa

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6,765	1	6,765	46,615	,000[a]
	Residual	2,612	18	,145		
	Total	9,378	19			

a. Predictors: (Constant), Tempo_orientação_manejo

b. Dependent Variable: Duraçao_AME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-,168	,396		-,424	,677		
	Tempo_orientação_manejo	,041	,006	,849	6,828	,000	1,000	1,000

a. Dependent Variable: Duraçao_AME

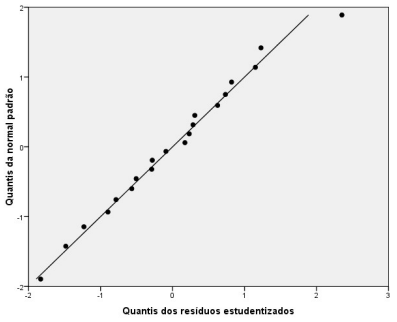
Modelo final : $\hat{Y}_i = -0{,}168 + 0{,}041\, X_i \;;\quad i = 1{,}2{,}...{,}\, 20$

$R^2 = 72{,}1\%$

66

Exemplo: Análise dos resíduos para o modelo 4

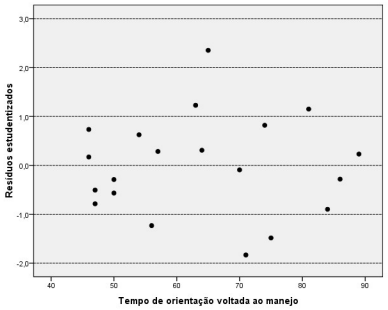
Figura 1: QQ-Plot (normalidade) para os resíduos estudentizados do modelo



67

Exemplo: Análise dos resíduos para o modelo 4

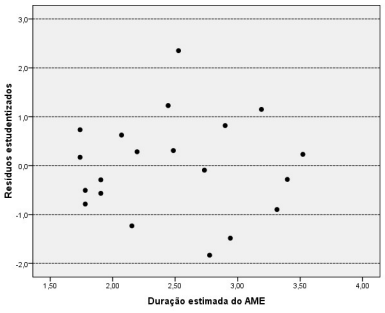
Figura 2: Gráfico de Dispersão entre o tempo de orientação voltada ao manejo e os resíduos estudentizados do modelo.



68

Exemplo: Análise dos resíduos para o modelo 4

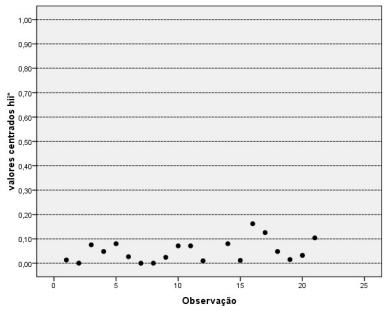
Figura 3: Gráfico de Dispersão entre a duração estimada do AME e os resíduos estudentizados do modelo.



69

Exemplo: Análise dos resíduos para o modelo 4

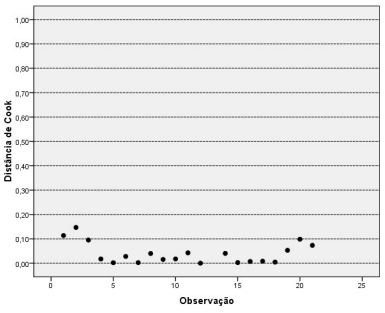
Figura 4: Gráfico de Dispersão entre a ordem das observações e as medidas centradas de alavancagem h_{ii}*



70

Exemplo: Análise dos resíduos para o modelo 4

Figura 5: Gráfico de Dispersão entre a ordem das observações e as medidas da distância de Cook



71

Aula prática - Exercício 1: Modelo de RLM com p-1=2 vars explicativas

Os dados apresentados na tabela a seguir se referem a um estudo sobre o desempenho de n=20 alunos de doutorado na 2ª avaliação da disciplina "Epidemiologia Social" (Y).

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

- *Tempo total de estudo = Tempo (em horas) dedicado a resolução de exercícios + Tempo (em horas) dedicado ao estudo da teoria.*
- *Nota da 1ª avaliação da referida disciplina.*

72

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

Aluno	Nota da V2	Tempo_exs	Tempo_teoria	Nota da V1
1	5,4	17	36	5,0
2	2,4	10	21	3,0
3	4,9	13	29	4,0
4	3,2	11	26	4,0
5	8,2	23	38	7,0
6	4,2	16	35	5,0
7	5,6	15	32	5,0
8	2,1	8	20	1,0
9	2,8	10	25	2,0
10	3,6	12	30	2,0
11	6,9	20	40	7,0
12	3,9	12	31	3,5
13	2,3	8	23	2,0
14	3,5	8	22	4,0
15	9,0	20	40	9,0
16	8,5	24	39	8,0
17	10,0	26	42	9,5
18	9,2	25	40	9,0
19	9,0	25	40	9,0
20	8,2	23	36	8,0

73

Aula prática - Exercício 1: Modelo de RLM com p-1=2 vars explicativas

a) Ajuste e selecione um modelo de regressão linear para dados observados e obtenha os resíduos estudentizados, as medidas (centradas) de alavancagem e os valores da distância de Cook (D).

b) Identifique a presença de outliers e de observações influentes no conjunto de dados. Avalie as hipóteses de homocedasticidade e normalidade dos erros.

c) Avalie mais detalhadamente a(s) observação(ões) com maior(es) valor(es) de D_i, isto é, elimine-a(s) e ajuste um novo modelo para os dados restantes. Verifique se as hipóteses de homocedasticidade e normalidade dos erros estão satisfeitas? Escreva a equação do modelo escolhido.

74

Aula prática - Exercício 2: Modelo de RLM com p-1=4 vars explicativas (“*Banco de hospitais”*)

Usando os dados do estudo com pacientes de n=113 hospitais, cujo objetivo é estudar a variação dos percentuais de pacientes com infecção hospitalar, avalie a *presença de observações influentes* no conjunto de dados, e tome as medidas necessárias. Explique passo a passo a suas decisões até a escolha do modelo final.

75