Modelos Lineares I

Regressão Linear Múltipla (RLM): Multicolinearidade

(27ª e 29ª Aulas)

26ª Aula → Vista da 1ª VE e do 1º trabalho (7/10/19)

28ª Aula → Aula de dúvidas para a 2ª VE (11/10/19)

Professor: Dr. José Rodrigo de Moraes

Universidade Federal Fluminense (UFF)

Departamento de Estatística (GET)

1

Modelo de Regressão Linear Múltipla (RLM):

Introdução:

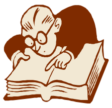
<u>Multicolinearidade</u>:

Quando as variáveis explicativas estão fortemente correlacionadas entre si (existência de multicolinearidade), a análise de regressão linear pode ficar confusa e desprovida de significado. Na presença de multicolinearidade, há dificuldade em distinguir o efeito de uma ou outra variável explicativa sobre a variável resposta Y.

- Multicolinearidade → passou a designar alta correlação (mas não necessariamente |1|) ≠ Multicolinearidade perfeita → combinação linear perfeita entre as variáveis.

2

Modelo de Regressão Linear Múltipla (RLM):



Introdução:

<u>Multicolinearidade</u>:

- Na presença de multicolinearidade, os parâmetros do modelo referentes as variáveis explicativas (multi)colineares podem ser altamente instáveis apresentando variâncias muito elevadas ou sinais inconsistentes.
- O problema de multicolinearidade não altera a qualidade global do ajustamento, pois não afeta a minimização da SQErros (a não ser que exista multicolinearidade perfeita !!!).

3

Modelo de Regressão Linear Múltipla (RLM)

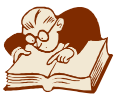
<u>Fontes de Multicolinearidade</u>:

Na prática, quase sempre existe algum tipo de correlação entre, pelo menos, algumas das variáveis independentes:

- Correlações meramente artificiais: duas variáveis que medem, por ex., o peso (uma em g, e a outra em kg);
- Correlações por uma variável ser função de outra (Ex.: Peso e IMC);

4

Modelo de Regressão Linear Múltipla (RLM):



<u>Indícios de Multicolinearidade</u>:

- Teste F de significância geral significativo (ou um R^2 alto) e os testes T pouco significativos (ou até mesmo não significativos);
- Sinais das estimativas pontuais dos parâmetros diferentes dos esperados.

5

Modelo de Regressão Linear Múltipla (RLM):

<u>Diagnóstico de Multicolinearidade</u>:

- Quando lidamos com mais de uma variável independente, pode-se representar a correlação linear entre cada possível par de variáveis, através da ***matriz de correlações bivariadas***.
- No caso de 4 variáveis independentes X_1, X_2, X_3 e X_4, existem $C_{4,2}$=6 correlações possíveis.

$$
\begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array}
\begin{array}{c}
\begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\
\left[\begin{array}{cccc} 1 & r_{12} & r_{13} & r_{14} \\ & 1 & r_{23} & r_{24} \\ & & 1 & r_{34} \\ & & & 1 \end{array}\right]
\end{array}
$$

6

Modelo de Regressão Linear Múltipla (RLM):

<u>Diagnóstico de Multicolinearidade</u>:

- Obter uma matriz de correlações bivariadas, onde existe |R| > 0,90 (*Atenção !!!*).

Observações:

- Não existe um valor de correlação limite a partir do qual seja possível identificar problemas na estimação do modelo, mas em geral utiliza-se o ponto 0,90.
- Estes *coeficientes de correlação linear de Pearson* (R) são válidos apenas para avaliar a presença de colinearidade entre variáveis duas-a-duas.

7

Modelo de Regressão Linear Múltipla (RLM):
<u>Diagnóstico de Multicolinearidade</u>:

- A medida estatística utilizada para diagnóstico de multicolinearidade é denominada de "<u>Fator de inflação da variância</u>", sendo denotada por *VIF* ("*Variance Inflation Factor*").
- A variância estimada de cada um dos estimadores dos parâmetros referentes às variáveis explicativas pode ser expressa por:

$$V\hat{A}R\left(\hat{\beta}_k\right) = \hat{\sigma}^2\left(\frac{1}{1-R_k^2}\right)\cdot\frac{1}{\sum_{i=1}^{n}\left(X_{ik}-\overline{X}_k\right)^2}, \quad \text{onde :} \quad VIF_k = \frac{1}{1-R_k^2} = \frac{1}{T_k}$$

Tolerância da variavel X_k

8

Modelo de Regressão Linear Múltipla (RLM):

<u>Diagnóstico de Multicolinearidade</u>:

$R_k^2 \rightarrow$ Coeficiente de determinação do modelo entre a variável X_k (como variável dependente) e as demais variáveis X_h como independentes ($h \neq k = 1,2,..., p-1$).

$T_k = 1 - R_k^2 \rightarrow$ Tolerância da variável X_k:

- $T_k \cong 0 \rightarrow$ a variável X_k representa uma combinação linear quase perfeita das demais variáveis (<u>instabilidade do parâmetro associado</u>).

$VIF_k = \dfrac{1}{1-R_k^2} \rightarrow$ Fator de inflação da variância para o parâmetro do modelo referente a variável X_k:

- *O VIF_k representa o aumento multiplicativo na estimativa da variância do estimador de β_k, devido a correlação de X_k com as demais variáveis independentes.*

9

Modelo de Regressão Linear Múltipla (RLM):

<u>Medidas de Diagnóstico de Multicolinearidade</u>:

Relações:

Se $R_k^2 \uparrow$ (multicolinearidade $\uparrow$) $\rightarrow T_k \downarrow \rightarrow VIF_k \uparrow \rightarrow V\hat{A}R\left(\hat{\beta}_k\right)\uparrow$
Se $R_k^2 \downarrow$ (multicolinearidade $\downarrow$) $\rightarrow T_k \uparrow \rightarrow VIF_k \downarrow \rightarrow V\hat{A}R\left(\hat{\beta}_k\right)\downarrow$

OBS :
Se $R_k^2 = 0$ (ausência de multicolinearidade) $\rightarrow T_k = 1 \rightarrow VIF_k = 1 \rightarrow \min\left[V\hat{A}R\left(\hat{\beta}_k\right)\right]$

- Variáveis com baixa tolerância T, tem valores elevados de *VIF*, e vice-versa.
- **Regra prática:**

VIF>10→R^2>0,90→R>0,95: diagnóstico positivo de multicolinearidade !!!

A variância infla por um fator superior a 10.

10

Exemplo 1: Modelo de RLM com p-1=5 vars explicativas

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).
- ✓ Tempo de orientação (em min) voltada a nutrição (X_3).
- ✓ Horas de trabalho (X_4)
- ✓ Tempo (em anos) de casada (X_5)

11

- **Exemplo 1 (*continuação*):**

Ajuste o modelo completo e verifique a existência de multicolinearidade. Em caso afirmativo, tome alguma medida para contornar este problema. Justifique o uso da medida escolhida.

Avalie as hipótese básicas do modelo, utilizando os resíduos estudentizados.

Escreva um pequeno relatório, com os resultados dos ajustes dos modelos e as principais conclusões obtidas (contexto do problema).

12

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

Mãe	AME	Anos de estudo	Tempo de orientação (nutrição)	Tempo de orientação (manejo)	Horas de trabalho	Tempo de casada
1	2,1	6	30	71	23	1
2	3,4	6	21	65	35	6
3	3,6	6	30	81	24	7
4	1,7	6	18	50	27	1
5	1,8	6	14	46	49	1
6	3,2	6	28	74	50	8
7	2,6	6	24	64	52	5
8	2,9	6	21	63	64	7
9	2,3	6	16	54	37	5
10	1,6	5	14	47	49	1
11	1,5	6	16	47	85	3
12	2,7	6	26	70	80	9
13	4,1	6	54	117	74	12
14	2,0	6	12	46	76	5
15	2,3	6	19	57	81	6
16	3,6	6	39	89	60	10
17	3,3	6	39	86	52	12
18	1,8	6	18	50	64	6
19	1,7	6	22	56	18	5
20	2,4	5	34	75	32	9
21	3,0	6	38	84	64	6

13

Exemplo: Resultados do ajuste do modelo completo (modelo 1)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11,291	5	2,258	56,329	,000[a]
	Residual	,601	15	,040		
	Total	11,892	20			

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

b. Dependent Variable: Duraçao_AME

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-2,298	,786		-2,924	,010
	Anos_estudo	,206	,153	,090	1,345	,199
	Tempo_orientação_nutrição	-,128	,022	-1,779	-5,854	,000
	Tempo_orientação_manejo	,102	,014	2,400	7,225	,000
	Horas_trabalho	-,003	,002	-,094	-1,416	,177
	Tempo_casado	,058	,023	,254	2,500	,025

a. Dependent Variable: Duraçao_AME

14

Exemplo: Resultados do ajuste do modelo completo (modelo 1)

Matriz de correlações bivariadas

Correlations

		Duraçao_AME	Anos_estudo	Tempo_orientação_nutrição	Tempo_orientação_manejo	Horas_trabalho	Tempo_casado
Pearson Correlation	Duraçao_AME	1,000	,394	,774	,874	,069	,782
	Anos_estudo	,394	1,000	,211	,275	,261	,170
	Tempo_orientação_nutrição	,774	,211	1,000	,979	,021	,731
	Tempo_orientação_manejo	,874	,275	,979	1,000	,044	,766
	Horas_trabalho	,069	,261	,021	,044	1,000	,283
	Tempo_casado	,782	,170	,731	,766	,283	1,000
Sig. (1-tailed)	Duraçao_AME	.	,038	,000	,000	,383	,000
	Anos_estudo	,038	.	,179	,114	,127	,231
	Tempo_orientação_nutrição	,000	,179	.	,000	,464	,000
	Tempo_orientação_manejo	,000	,114	,000	.	,426	,000
	Horas_trabalho	,383	,127	,464	,426	.	,107
	Tempo_casado	,000	,231	,000	,000	,107	.
N	Duraçao_AME	21	21	21	21	21	21
	Anos_estudo	21	21	21	21	21	21
	Tempo_orientação_nutrição	21	21	21	21	21	21
	Tempo_orientação_manejo	21	21	21	21	21	21
	Horas_trabalho	21	21	21	21	21	21
	Tempo_casado	21	21	21	21	21	21

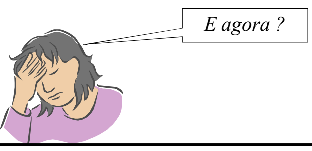
15

Exemplo: Resultados do ajuste do modelo completo (modelo 1) – Valores de ***VIF***

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-2,298	,786		-2,924	,010		
	Anos_estudo	,206	,153	,090	1,345	,199	,751	1,332
	Tempo_orientação_nutrição	-,128	,022	-1,779	-5,854	,000	,037	27,377
	Tempo_orientação_manejo	,102	,014	2,400	7,225	,000	,031	32,735
	Horas_trabalho	-,003	,002	-,094	-1,416	,177	,764	1,309
	Tempo_casado	,058	,023	,254	2,500	,025	,327	3,059

a. Dependent Variable: Duraçao_AME



16

Exemplo: Resultados do ajuste do modelo 2

Modelo 2: Excluindo "*Tempo de orientação voltada a nutrição*"

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 B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics 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

17

Exemplo: Resultados do ajuste do modelo 3

Modelo 3: Excluindo "*Horas de trabalho*"

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 B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics 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

18

Exemplo: Resultados do ajuste do modelo 4

Modelo 4: Excluindo "*Anos de estudo*"

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	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			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

19

Exemplo: Resultados do ajuste do modelo 5

Modelo 5: Excluindo "*Tempo de casado*"

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

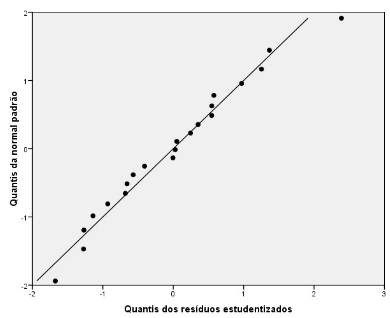
Modelo final : $\hat{Y}_i = 0{,}099 + 0{,}037\,X_i \; ; \quad i = 1,2,\dots,21$

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

20

Exemplo: Análise dos resíduos para o modelo 5

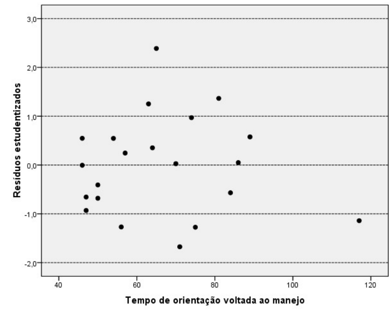
Figura 1: QQ-Plot (normalidade) para os resíduos estudentizados do modelo



21

Exemplo: Análise dos resíduos para o modelo 5

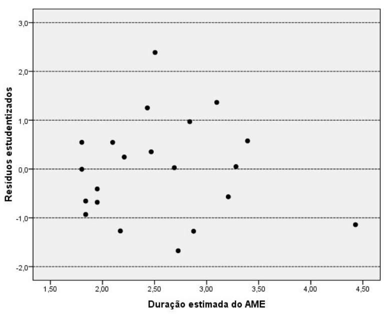
Figura 2: Gráfico de Dispersão entre o tempo de orientação voltada ao manejo e os resíduos estudentizados do modelo.



22

Exemplo: Análise dos resíduos para o modelo 5

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



23

Exemplo 2: Modelo de RLM com p-1=3 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 (em horas) dedicado a resolução de exercícios (X_1)
- ✓ Tempo (em horas) dedicado ao estudo da teoria (X_2).
- ✓ Nota da 1ª avaliação da referida disciplina (X_3).

24

❑ **Exemplo 2 (*continuação*):**

Ajuste o modelo completo e verifique a existência de multicolinearidade. Em caso afirmativo, tome alguma medida para contornar este problema. Justifique o uso da medida escolhida.

Avalie 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.

Escreva um pequeno relatório, com os resultados dos ajustes dos modelos e as principais conclusões obtidas (contexto do problema). OBS: *Ajuste usando o SPSS !!!*

25

Banco de Dados: *Modelo de RLM (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

26

Exemplo: Resultados do ajuste do modelo completo (modelo 1)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,984[a]	,968	,962	,53113

a. Predictors: (Constant), Tempo_teoria, nota_v1, Tempo_exs

b. Dependent Variable: Y_nota_v2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	135,276	3	45,092	159,844	,000[a]
	Residual	4,514	16	,282		
	Total	139,790	19			

a. Predictors: (Constant), Tempo_teoria, nota_v1, Tempo_exs

b. Dependent Variable: Y_nota_v2

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)	-,625	,819		-,763	,457	-2,361	1,111
	nota_v1	,514	,135	,521	3,794	,002	,227	,801
	Tempo_exs	,182	,075	,433	2,442	,027	,024	,341
	Tempo_teoria	,017	,049	,046	,351	,730	-,086	,120

27

a. Dependent Variable: Y_nota_v2

Exemplo: Resultados do ajuste do modelo completo (modelo 1)

Matriz de correlações bivariadas

Correlations

		Y_nota_v2	nota_v1	Tempo_exs	Tempo_teoria
Pearson Correlation	Y_nota_v2	1,000	,971	,968	,920
	nota_v1	,971	1,000	,945	,897
	Tempo_exs	,968	,945	1,000	,939
	Tempo_teoria	,920	,897	,939	1,000
Sig. (1-tailed)	Y_nota_v2	.	,000	,000	,000
	nota_v1	,000	.	,000	,000
	Tempo_exs	,000	,000	.	,000
	Tempo_teoria	,000	,000	,000	.
N	Y_nota_v2	20	20	20	20
	nota_v1	20	20	20	20
	Tempo_exs	20	20	20	20
	Tempo_teoria	20	20	20	20

28

Exemplo: Resultados do ajuste do modelo completo (modelo 1) – Valores de *VIF*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-,625	,819		-,763	,457		
	nota_v1	,514	,135	,521	3,794	,002	,107	9,357
	Tempo_exs	,182	,075	,433	2,442	,027	,064	15,547
	Tempo_teoria	,017	,049	,046	,351	,730	,117	8,555

a. Dependent Variable: Y_nota_v2

29

Exemplo: Resultados do ajuste do modelo 2

Modelo 2: Criando "*Tempo total de estudo (Exs + Teoria)*"

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	134,660	2	67,330	223,150	,000[a]
	Residual	5,129	17	,302		
	Total	139,790	19			

a. Predictors: (Constant), Tempo_total_estudo, nota_v1

b. Dependent Variable: Y_nota_v2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-1,352	,677		-1,996	,062		
	nota_v1	,596	,128	,605	4,674	,000	,129	7,764
	Tempo_total_estudo	,078	,026	,392	3,031	,008	,129	7,764

a. Dependent Variable: Y_nota_v2

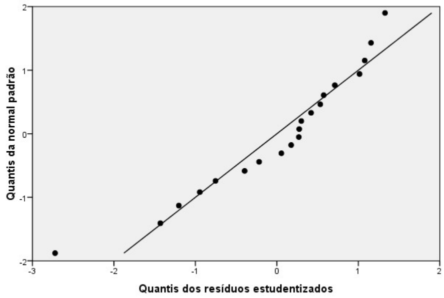
Modelo final : $\hat{Y}_i = -1{,}352 + 0{,}596\, X_{1i} + 0{,}078\, X_{2i}$; $i = 1,2,\dots, 20$

$R^2 = 96{,}3\%$

30

Exemplo: Análise dos resíduos para o modelo 2

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



31

Exemplo: Análise dos resíduos para o modelo 2

Teste de normalidade de *Kolmogorov-Smirnov* para os resíduos do modelo

One-Sample Kolmogorov-Smirnov Test

		Unstandardized Residual	Standardized Residual	Studentized Residual
N		20	20	20
Normal Parameters[a,b]	Mean	,0000000	,0000000	,0110334
	Std. Deviation	,51958174	,94590530	1,01117996
Most Extreme Differences	Absolute	,172	,172	,168
	Positive	,102	,102	,096
	Negative	-,172	-,172	-,168
Kolmogorov-Smirnov Z		,768	,768	,749
Asymp. Sig. (2-tailed)		,597	,597	,629

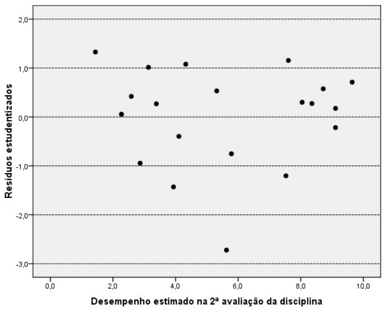
a. Test distribution is Normal.

b. Calculated from data.

32

Exemplo: Análise dos resíduos para o modelo 2

Figura 2: Gráfico de Dispersão entre o desempenho estimado na 2ª avaliação da disciplina e os resíduos estudentizados do modelo.



33

Aula prática – Exercício 1: Modelo de Regressão Linear Múltipla (RLM)

Considerando os dados sobre o número de publicações nos últimos 5 anos (X_1), o tempo de experiência (em meses) como docente (X_2) e o índice de desempenho como professor/pesquisador (numa escala de 0 a 100), pede-se:

a) Construa um gráfico de dispersão entre as duas variáveis explicativas X_1 e X_2, e calcule o coeficiente de correlação linear entre elas.Você acha que existe problema de multicolinearidade?

b) Ajuste o modelo com as duas variáveis explicativas, e avalie se sua conclusão obtida na letra a) continua a mesma.

c) Obtenha o "fator de inflação da variância" (*VIF*) para cada uma das variáveis explicativas. Qual a sua conclusão ?

d) Afinal qual modelo você escolheria? Sugestão: *Use o teste de comparabilidade de modelos e/ou alguma medida apropriada.*

34

Tabela 1: ***Dados sobre n=14 profs. de uma universidade***

Prof.	X_1	X_2	Y
1	17	35,7	52,1
2	6	11,4	24,6
3	13	28,6	49,2
4	11	25,8	30,0
5	23	50,6	82,2
6	16	27,2	42,4
7	15	31,3	55,7
8	5	10,0	21,1
9	10	18,9	27,7
10	12	25,2	36,3
11	20	39,9	69,1
12	12	32,5	38,8
13	8	13,6	22,8
14	8	19,0	34,7

35

Aula prática – Exercício 2: Modelo de Regressão Linear Múltipla (RLM)

Considerando os dados sobre o salário (em 100 UM), idade (em anos) e tempo de serviço (em anos) de n=25 trabalhadores de uma pequena empresa (Tabela 2), pede-se:

a) Construa um gráfico de dispersão entre as duas variáveis explicativas.

b) Calcule o coeficiente de correlação linear entre as variáveis explicativas. Você acha que existe problema de multicolinearidade?

c) Ajuste o modelo com as duas variáveis explicativas, e avalie se sua conclusão obtida na letra b) continua a mesma.

d) Obtenha o "fator de inflação da variância" (*VIF*) para cada uma das variáveis explicativas.

e) Qual a sua conclusão final ?

36

Tabela 2: ***Dados sobre n=25 funcionários de uma empresa***

continuação

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

37