Modelos Lineares I

Regressão Linear Múltipla (RLM):

Heterocedasticidade

(32ª, 33ª e 34ª Aulas)

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

1

Modelo de Regressão Linear:

Introdução:

Heterocedasticidade:

Representa uma das violações básicas do modelo que se pode detectar por meio da análise gráfica dos resíduos ou de outros métodos formais.

Heterocedasticidade → variância dos erros não é constante.

***Conseqüências*:**

- Os estimadores de MQO continuam não viciados, mas não são mais os melhores estimadores lineares não viciados;
- As variâncias dos estimadores de MQO são incorretos, invalidando as inferências.

2

Modelo de Regressão Linear:

Introdução:

Fontes de Heterocedasticidade:

- Uso de dados de médias;
- Diferentes observadores;
- Valores discrepantes;
- Natureza das variáveis.

3

Modelo de Regressão Linear:

Introdução:

Idéia: Pesquisa para estudar o salários em função dos anos de estudo, isto é, Salário = *f* (anos de estudo)

Se você tivesse que estimar os salários para "*pessoas com baixa escolaridade*" e para "*pessoas com alta escolaridade*".

Pergunta: *Para que "grupo" você acha que seria mais fácil de estimar (ou para que "grupo" seria mais fácil fazer suposições sobre os seus salários) ?*

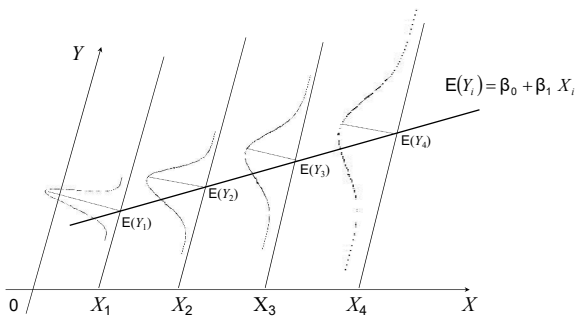
***Comentários*:**

- **Média:** relação crescente entre os "anos de estudo" e "salário";
- **Variância:** Pessoas menos escolarizadas → $\sigma_Y^2 \downarrow$
  Pessoas mais escolarizadas → $\sigma_Y^2 \uparrow$

Outro exempo: Poupança = *f*(renda)

4

Existência de heterocedasticidade:



5

Exemplo 1: Modelo de Regressão Linear

Os dados apresentados na tabela a seguir se referem a um estudo sobre os salários (em UM) e a escolaridade (em anos de estudo) para uma amostra de n=20 trabalhadores de uma indústria. Pede-se:

a) Faça o gráfico de dispersão entre X e Y.

b) Ajuste o modelo completo e verifique a existência ou não de heterocedasticidade, utilizando a análise gráfica dos resíduos estudentizados.

6

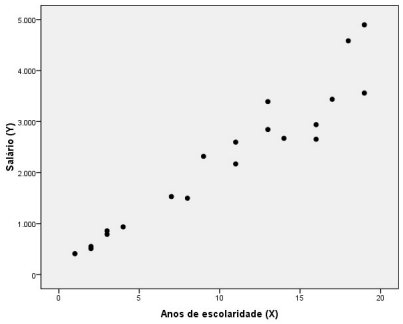
Banco de Dados: *Modelo de RLS (n=20 trabalhadores)*:

Trabalhador	Anos de estudo	Salário
1	1	410,00
2	2	508,90
3	3	857,70
4	2	551,30
5	3	789,20
6	4	935,50
7	7	1.529,30
8	8	1.497,50
9	9	2.317,70
10	11	2.169,50
11	11	2.596,80
12	13	2.844,60
13	13	3.391,00
14	14	2.671,20
15	16	2.653,80
16	16	2.939,10
17	17	3.437,00
18	18	4.583,30
19	19	3.559,30
20	19	4.896,70

7

Exemplo 1: Análise gráfica (n=20 trabalhadores)

Figura 1: Gráfico de dispersão entre os anos de estudo (X) e salário (Y)



8

Exemplo 1: Resultados do ajuste do modelo original

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,953[a]	,908	,903	418,41013

a. Predictors: (Constant), anos_escol

b. Dependent Variable: salário

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,121E7	1	3,121E7	178,281	,000[a]
	Residual	3151206,674	18	175067,037		
	Total	3,436E7	19			

a. Predictors: (Constant), anos_escol

b. Dependent Variable: salário

9

Exemplo 1: Resultados do ajuste do modelo original

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	139,074	184,155		,755	,460
	anos_escol	205,621	15,400	,953	13,352	,000

a. Dependent Variable: salário

Modelo original :

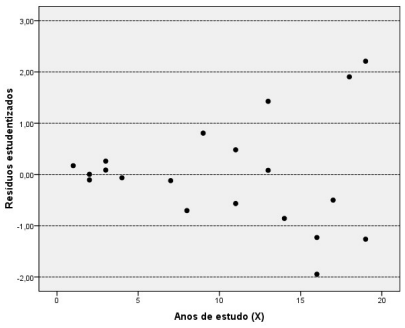
$$\hat{Y}_i = 139{,}074 + 205{,}621\,X_i \;;\quad i = 1,2,...,20$$

$$R^2 = 31.211.075{,}428/34.362.282{,}102 \cong 90{,}8\%$$

10

Exemplo 1: Análise dos resíduos do modelo original

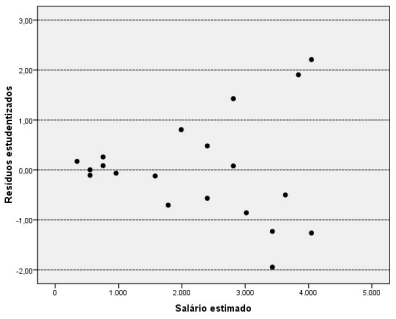
Figura 2: Gráfico de dispersão entre os anos de estudo e os resíduos estudentizados do modelo.



11

Exemplo 1: Análise dos resíduos do modelo original

Figura 3: Gráfico de dispersão entre o salário estimado e os resíduos estudentizados do modelo.



12

I) Teste de White - Homocedasticidade:

A hipótese de homocedasticidade também pode ser avaliada pelo "*Teste de White*".

Ilustração: No caso do modelo de RLM com p-1=2 variáveis independentes:

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

- É feita uma regressão auxiliar em que a variável dependente é o resíduo bruto ao quadrado da regressão original:

$$e_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i1}^2 + \gamma_4 X_{i2}^2 + \gamma_5 X_{i1} X_{i2} + u_i$$

13

I) Teste de White - Homocedasticidade:

1) Hipóteses a serem testadas:

$$\begin{cases} H_0 : \text{Homocedasticidade} \\ H_1 : \text{Heterocedasticidade} \end{cases}$$

2) Estatística de Teste:

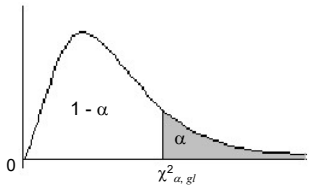
$$W = nR^2 \sim \chi^2_{gl} \quad ; \quad \text{onde :}$$

$gl \rightarrow$ nº de variáveis independentes na regressão auxiliar

14

I) Teste de White - Homocedasticidade:

3) Região crítica:



$$RC = \left\{ W \in \Re \,/\, w \geq \chi_{\alpha, gl} \right\}$$

15

I) Teste de White - Homocedasticidade:

4) Tomada de Decisão:

- Se $w_{obs} \in RC$ rejeita-se a hipótese nula "H_0: homocedasticidade" ao nível de significância α.
- Se $w_{obs} \notin RC$ não há evidências para rejeitar a hipótese nula "H_0: homocedasticidade" ao nível de significância α.

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

16

❑ **Exemplo 1: Anos de estudo *vs* Salário (n=20 trabs).**

Usando os dados sobre os anos de estudo e salários dos n=20 trabalhadores de uma indústria, aplique o "*Teste de White*" para verificar a existência de violação da hipótese básica do modelo (*heterocedasticidade*). Qual a sua conclusão?

Modelo original:

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

Pergunta: Qual a regressão auxiliar neste caso ?

17

Exemplo 1: Teste de White – Modelo original

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	38435,605	101913,265		,377	,711
	anos_escol	-19713,094	24903,833	-,547	-,792	,440
	anos_escol_quad	2252,934	1216,132	1,281	1,853	,081

a. Dependent Variable: Res_bruto_quad

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,758[a]	,575	,525	1,54773E5

a. Predictors: (Constant), anos_escol_quad, anos_escol

b. Dependent Variable: Res_bruto_quad

$$w_{obs} = nR^2 = 20 \cdot 0{,}575 = 11{,}5 > \chi^2_{0{,}05;2} = 5{,}991 \rightarrow$$

$\rightarrow$ Rejeita-se a hipótese de homocedasticidade

18

Modelo de Regressão Linear:

***Correção de Heterocedasticidade (proporcional)*:**

- Qual o padrão/causa da heterocedasticidade ?
- Em fenômenos sociais, econômicos e biológicos, em geral, a variância (ou desvio-padrão) dos erros são supostamente proporcionais a X, isto é:

$$VAR(\varepsilon_i) = \sigma_i^2 = \sigma^2 \cdot X_i^2 \quad \rightarrow \quad DP(\varepsilon_i) = \sigma \cdot X_i$$

19

Modelo de Regressão Linear

Correção de Heterocedasticidade (proporcional):

- Transformar uma variável (Y) cuja variância é $VAR(\varepsilon_i) = \sigma^2 X_i^2$ em outra variável (Y*) cuja variância é σ^2.

Demonstração:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\frac{Y_i}{X_i} = \beta_0 \frac{1}{X_i} + \beta_1 + \frac{\varepsilon_i}{X_i}$$

$$Y_i^* = \beta_1 + \beta_0 \; X_i^* + \varepsilon_i^*$$

Modelo transformado: $VAR(\varepsilon_i^*) = \sigma^2$

20

Exemplo 2 (*Continuação do Exemplo* 1):

Considerando ainda os dados sobre os salários (em UM) e a escolaridade (em anos de estudo) de n=20 trabalhadores de uma indústria:

a)Reajuste o modelo corrigindo o problema de heterocedasticidade.

b) Escreva a equação do modelo, e interprete as estimativas dos parâmetros do modelo.

c) Avalie também a hipótese de normalidade para este modelo.

d) Refaça o teste de White para o modelo transformado, a fim de checar se o problema de heterocedasticidade foi resolvido.

21

Exemplo 2 - a) Resultados do ajuste do modelo transformado

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,832[a]	,691	,674	30,70106

a. Predictors: (Constant), anos_escol_transf

b. Dependent Variable: Salário_transf

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	38025,775	1	38025,775	40,343	,000[a]
	Residual	16965,988	18	942,555		
	Total	54991,762	19			

a. Predictors: (Constant), anos_escol_transf

b. Dependent Variable: Salário_transf

22

Exemplo 2 - a) Resultados do ajuste do modelo transformado

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	198,869	9,126		21,791	,000
	anos_escol_transf	188,745	29,716	,832	6,352	,000

a. Dependent Variable: Salário_transf

Modelo transformado :

$$\hat{Y}_i^* = 198{,}869 + 188{,}745 \; X_i^* \quad ; \quad i = 1,2,...,20$$

23

Exemplo 2 – b) Anos de estudo (X*=1/X) *versus* Salário (Y*=Y/X).

Modelo transformado : $\hat{Y}_i^* = 198{,}869 + 188{,}745 \; X_i^* \quad ; \quad i = 1,2,...,20$

Em termos das variáveis originais :

$$\frac{\hat{Y}_i}{X_i} = 198{,}869 + 188{,}745 \frac{1}{X_i} \quad \rightarrow \quad \hat{Y}_i = 198{,}869 \, X_i + 188{,}745 \quad \rightarrow$$

$$\rightarrow \quad \boxed{\hat{Y}_i = 188{,}745 + 198{,}869 \, X_i \quad ; \quad i = 1,2,...,20}$$

24

Exemplo 2 – b) <u>Reconstruindo a tabela com as estimativas</u>:

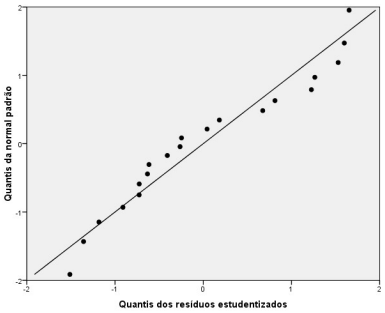
Resultados obtidos a partir do modelo transformado:

Variáveis originais	Estimativa pontual dos parâmetros	Medida de Precisão (DP)	Estatística de teste (T)	P-valor
Constante	188,745	29,716	6,352	<0,001
Anos de estudo (X)	198,869	9,126	21,791	<0,001

25

Exemplo 2-c): Hipótese de normalidade dos erros

Figura 3: *QQ-Plot* (normalidade) para os resíduos estudentizados (r^{S*}) do modelo transformado



26

Exemplo 2 – d) Teste de White – Modelo transformado:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1172,566	337,864		3,471	,003
	anos_escol_transf	-2546,962	2502,224	-,747	-1,018	,323
	anos_escol_transf_quad	2026,627	2615,736	,568	,775	,449

a. Dependent Variable: Res_bruto_quad_transf

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,275[a]	,076	-,033	821,86055

a. Predictors: (Constant), anos_escol_transf_quad, anos_escol_transf

b. Dependent Variable: Res_bruto_quad_transf

$w_{obs} = nR^2 = 20 \cdot 0{,}076 = 1{,}52 < \chi^2_{0{,}05;2} = 5{,}991 \rightarrow$

$\rightarrow$ Não rejeita-se a hipótese de homocedasticidade

27

Aula prática / Sala - Exercício 1 ("*Saídas*"):

A associação industrial de um determinada cidade tem como objetivo verificar se existe relação entre o "número de trabalhadores (X)" e o número de supervisores dos estabelecimentos associados (Y)".

A Tabela 1 fornece essas informações para uma amostra de 27 estabelecimentos.

28

Aula prática - Exercício 1 (*continuação*):

Tabela 1: Dados sobre n=20 estabelecimentos

Estab.	Nº de trab. (X)	Nº de superv. (Y)	Estab.	Nº de trab. (X)	Nº de superv. (Y)
			continuação		
1	294	30	15	615	100
2	247	32	16	999	109
3	267	37	17	1022	114
4	358	44	18	1015	117
5	423	47	19	700	106
6	311	49	20	850	128
7	450	56	21	980	130
8	534	62	22	1025	160
9	438	68	23	1021	97
10	697	78	24	1200	180
11	688	80	25	1250	112
12	630	84	26	1500	210
13	709	88	27	1650	135
14	627	97			

29

- **Aula prática – Exercício 1 (*continuação*):**

a) Ajuste o modelo e avalie a existência ou não de violação da hipótese básica do modelo (*heterocedasticidade*). Para tanto use análises gráficas (inclusive usando os resíduos estudentizados e o teste de White (*Defina as Hipóteses a serem testadas*, *Estatística de teste*, *Região Crítica* e *Tomada de decisão*). Qual a conclusão obtida?

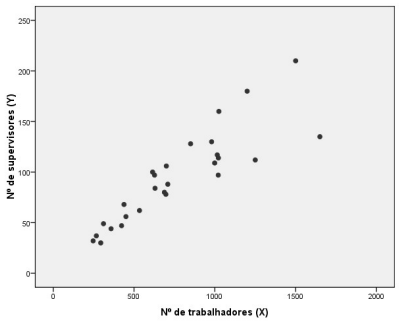
b) Caso necessário, corrija a heterocedasticidade, e analise as estimativas dos parâmetros do modelo.

c) Avalie as hipótese básicas do modelo.

30

Exercício 1 – a):

Figura 1: Gráfico de Dispersão entre o nº de trabalhadores (X) e o nº de supervisores (Y).



31

Exercício 1 –a): Resultados do ajuste do modelo (original) explicativo do nº de supervisores (Y), em função do nº de trabalhadores (X).

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,881[a]	,776	,767	21,72930

a. Predictors: (Constant), N_trab_X

b. Dependent Variable: N_superv_Y

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	40862,603	1	40862,603	86,544	,000[a]
	Residual	11804,064	25	472,163		
	Total	52666,667	26			

a. Predictors: (Constant), N_trab_X

b. Dependent Variable: N_superv_Y

32

Exercício 1 – a):

Resultados do ajuste do modelo (original) explicativo do nº de supervisores (Y), em função do nº de trabalhadores (X).

Coefficients[a]

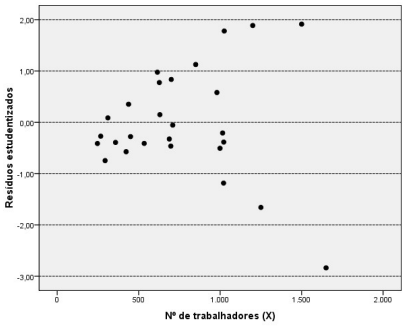
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	14,448	9,562		1,511	,143
	N_trab_X	,105	,011	,881	9,303	,000

a. Dependent Variable: N_superv_Y

33

Exercício 1 – a):

Figura 2: Gráfico de dispersão entre o nº de trabalhadores (X) e os resíduos estudentizados do modelo original.



34

Exercício 1 – a): Teste de White – modelo original

Regressão auxiliar

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	497,488	281,930		1,765	,090
	N_trab_X	-1,808	,718	-1,009	-2,518	,019
	N_trab_X_quad	,002	,000	1,832	4,571	,000

a. Dependent Variable: RES_quad

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,885[a]	,783	,765	326,97562

a. Predictors: (Constant), N_trab_X_quad, N_trab_X

35

Exercício 1 – b):

Resultados do ajuste do modelo (transformado) explicativo do nº de supervisores (Y*=Y/X), em função do nº de trabalhadores (X*=1/X).

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,164[a]	,027	-,012	,022664789

a. Predictors: (Constant), N_trab_X_transf

b. Dependent Variable: N_superv_Y_transf

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,000	1	,000	,693	,413[a]
	Residual	,013	25	,001		
	Total	,013	26			

a. Predictors: (Constant), N_trab_X_transf

b. Dependent Variable: N_superv_Y_transf

36

Exercício 1 – b):

Resultados do ajuste do modelo (transformado) explicativo do nº de supervisores (Y*=Y/X), em função do nº de trabalhadores (X*=1/X).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	,121	,009		13,445	,000
	N_trab_X_transf	3,803	4,570	,164	,832	,413

a. Dependent Variable: N_superv_Y_transf

37

Exercício 1 – b)

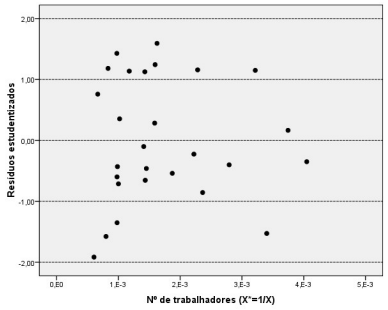
Preencha a tabela abaixo a partir dos resultados obtidos com o modelo transformado:

Variáveis originais	Estimativa pontual dos parâmetros	Medida de Precisão (DP)	Estatística de teste (T)	P-valor
Constante				
Nº de trab. (X)				

38

Exercício 1 – c): *Hipóteses básicas do modelo*

Figura 3: Gráfico de dispersão entre o nº de trabalhadores (X*=1/X) e os resíduos estudentizados (r^{S*}) do modelo transformado.



39

Exemplo 1 - c): *Hipóteses básicas do modelo*

Figura 4: *QQ-Plot* (normalidade) para os resíduos estudentizados (r^{S*}) do modelo transformado

40

Aula prática – Exercício 2: Voltando a um dos exemplos de aplicação: Modelo de RLM com p-1=2 variáveis explicativas

A tabela 1 a seguir fornece o valor dos *salários* (em 100 UM), a idade e o tempo de serviço 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), em anos.
- ✓ *Tempo de serviço* (X_2), em anos.

41

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				

42

❑ **Aula prática - Exercício 2 (*continuação*):**

Ajuste o modelo e verifique a existência de *heterocedasticidade* usando:

a) a análise gráfica dos resíduos brutos e estudentizados.

b) o teste de White (*Hipóteses a serem testadas*, *Estatística de teste*, *Região Crítica* e *Tomada de decisão*).

Qual a conclusão obtida ?

Analise as estimativas dos parâmetros do modelo (sem heterocedasticidade) e calcule e interprete o coeficiente de determinação do modelo.

Resp.: $w_{obs} = 4{,}975\ (gl = 5)$

43

Aula prática – Exercício 3:

Os dados apresentados na tabela 2 a seguir se referem a um estudo sobre o *consumo de energia elétrica* (kwh/mês) para uma amostra de 17 cidades.

O objetivo do estudo é estudar a relação o consumo de energia (Y) e as seguintes variáveis explicativas:

✓*Tarifa* (X_1), *em UM/kwh.*

✓*Renda mensal familiar* (X_2), em *UM/mês*

44

Tabela 2: ***Dados sobre n=17 cidades***

Cidade	Consumo (Y)	Tarifa (X_1)	Renda (X_2)
1	355,70	1,50	600
2	393,80	1,80	400
3	429,10	2,00	700
4	250,50	1,20	300
5	484,90	1,30	600
6	377,10	1,60	700
7	194,30	3,00	500
8	328,20	2,50	600
9	498,60	2,20	850
10	444,50	1,90	550
11	217,10	0,90	300
12	279,80	1,10	700
13	300,90	1,50	800
14	199,80	1,40	650
15	798,20	1,30	900
16	483,40	1,80	500
17	518,90	2,40	400

45

❑ **Aula prática - Exercício 3 (*continuação*): Consumo de energia elétrica em n=17 cidades.**

a)Ajuste o modelo completo usando o *programa RStudio* (com as duas variáveis explicativas) e avalie a significância dos parâmetros do modelo (e o sentido das associações), fixando o nível de significância de 10%.

b) Avalie se existe violação da hipótese de homocedasticidade para o modelo ajustado na letra a), usando a análise gráfica dos resíduos estudentizados e o teste de White. Qual a sua conclusão ?

c) Exclua a variável com efeito não significativo ao nível de 10%, e ajuste um novo modelo. Avalie se o problema de heterocedasticidade permanece ou não neste modelo.

46

❑ **Aula prática - Exercício 3 (*continuação*): Consumo de energia elétrica em n=17 cidades.**

d) Escreva a equação do modelo ajustado na letra c), e interprete as estimativas do parâmetro do modelo e o coeficiente de determinação do modelo.

e) Construa o *QQ Plot* para os resíduos estudentizados.

f) O modelo é apropriado?

OBS: *Use o programa R e/ou SPSS.*

Resp.: b) $w_{obs} = 13{,}498\ (gl = 5)$

c) $w_{obs} = 3{,}927\ (gl = 2)$

47

Avisos:

✓ Fazer a **4ª Lista de Exercícios** da disciplina Modelos Lineares I ("Análise de Regressão") proposta pelo Prof. *Dr. José Rodrigo de Moraes (GET/UFF)*.

48