Modelos Lineares I

Regressão Linear Múltipla (RLM):

Inferência no modelo e Análise dos resíduos

(21ª e 22ª Aulas)

Professor: Dr. José Rodrigo de Moraes

Universidade Federal Fluminense (UFF)

Departamento de Estatística (GET)

1

Modelo de Regressão Linear Múltipla:

Como vimos:

$$\hat{\boldsymbol{\beta}} \sim N\left[\boldsymbol{\beta}, \sigma^2 (X`X)^{-1}\right]$$

ou seja, o vetor $\hat{\boldsymbol{\beta}}$ tem distribuição normal com:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad \text{e} \quad VAR(\hat{\boldsymbol{\beta}}) = \sigma^2 (X`X)^{-1}$$

A estimativa da matriz de variância-covariância é dada por;

$$\hat{VAR}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (X`X)^{-1}, \quad \text{onde:}$$

$$\hat{\sigma}^2 = QMRes = \frac{SQRes}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

2

Modelo de Regressão Linear Múltipla:

Seja C_{kk} o k-ésimo elemento da diagonal principal da matriz $(X`X)^{-1}$. Sabe-se que a variância do estimador de β_k, k=0,1,2,...,p-1 é dada por:

$$VAR(\hat{\beta}_k) = \sigma^2 C_{kk} \quad \rightarrow \quad DP(\hat{\beta}_k) = \sigma\sqrt{C_{kk}}$$

Como:

$$Z = \frac{\hat{\beta}_k - \beta_k}{\sigma\sqrt{C_{kk}}} \sim N(0,1) \quad \text{e} \quad \chi^2 = \frac{(n-p)QMRes}{\sigma^2} \sim \chi^2_{n-p},$$

temos que:

$$T = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{VAR}(\hat{\beta}_k)}} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{QMRes}\sqrt{C_{kk}}} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{QMRes \cdot C_{kk}}} \sim T\text{-Student com } (n-p) \text{ g.l`s}$$

3

Teste de Significância Individual para o parâmetro β_k:

- **Hipóteses a serem testadas:**

$$\begin{cases} H_0: \beta_k = 0 \\ H_1: \beta_k \neq 0 \end{cases}$$

- **Estatística de Teste:**

$$T = \frac{\hat{\beta}_k}{\sqrt{\hat{VAR}(\hat{\beta}_k)}} \sim T_{n-p}$$

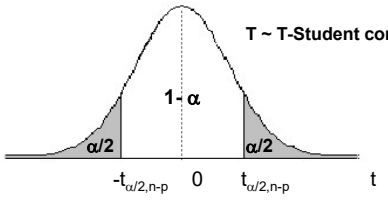
$$t_{obs} = \frac{\hat{\beta}_k}{\sqrt{\hat{VAR}(\hat{\beta}_k)}}$$

Valor observado de T

4

Teste de Significância Individual para o parâmetro β_k:

- **Região crítica:**



$$RC = \left\{ t \in \Re \ / \ t \leq -t_{\alpha/2,n-p} \text{ ou } t \geq t_{\alpha/2,n-p} \right\}$$

ou equivalentemente:

$$RC = \left\{ t \in \Re \ / \ |t| \geq t_{\alpha/2,n-p} \right\}$$

5

Teste de Significância Individual para o parâmetro β_k:

- **Tomada de Decisão:**

- Se $t_{obs} \in RC$ rejeita-se $H_0: \beta_k = 0$ ao nível de significância α, e conclui-se que existe relação linear significativa entre X_k e Y.
- Se $t_{obs} \notin RC$ não há evidências para rejeitar $H_0: \beta_k = 0$ ao nível de significância α, e conclui-se que não existe relação linear significativa entre X_k e Y.

OBS: Ou então, usar o método do p-valor:

p-valor $\leq \alpha$=0,05 → rejeita-se H_0 ao nível de 5%.

6

Intervalo de Confiança para o parâmetro β_k :

- Para construir um intervalo de confiança (IC) para β_k ao nível de confiança de 100(1-α)%, calcula-se a probabilidade:

$$P\left[-t_{\alpha/2,n-p} \le \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{VAR}(\hat{\beta}_k)}} \le t_{\alpha/2,n-p}\right] = 1-\alpha$$

$$P\left[\hat{\beta}_k - t_{\alpha/2,n-p}\sqrt{\hat{VAR}(\hat{\beta}_k)} \le \beta_k \le \hat{\beta}_k + t_{\alpha/2,n-p}\sqrt{\hat{VAR}(\hat{\beta}_k)}\right] = 1-\alpha$$

onde:

$$\hat{DP}(\hat{\beta}_k) = \sqrt{\hat{VAR}(\hat{\beta}_k)}$$

Limite superior (l_{sup}) do intervalo

Limite inferior (l_{inf}) do intervalo

7

Voltando ao exemplo de aplicação: Modelo de Regressão Linear Múltipla com p-1=2 variáveis explicativas.

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)

8

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				

9

Resultados do Ajuste do Modelo – *RLM Normal*:

Analyse / Regression / Linear

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	12,472	2,710		4,603	,000
	Idade	,304	,106	,459	2,868	,009
	Tempo_serv	,474	,154	,493	3,078	,006

a. Dependent Variable: Salário_Y

Estimativas dos parâmetros

10

Resultados do Ajuste do Modelo – *RLM Normal*:

Analyse / Regression / Linear

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	12,472	2,710		4,603	,000
	Idade	,304	,106	,459	2,868	,009
	Tempo_serv	,474	,154	,493	3,078	,006

a. Dependent Variable: Salário_Y

Estimativas dos erros-padrões

11

Matriz de variância-covariância estimada para os estimadores dos parâmetros:

Analyse /Regression / Linear

Coefficient Correlations[a]

Model			Tempo_serv	Idade
1	Correlations	Tempo_serv	1,000	-,838
		Idade	-,838	1,000
	Covariances	Tempo_serv	,024	-,014
		Idade	-,014	,011

a. Dependent Variable: Salário_Y

12

Resultados do Ajuste do Modelo – *RLM Normal*:

Analyse / Regression / Linear

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	12,472	2,710		4,603	,000
	Idade	,304	,106	,459	2,868	,009
	Tempo_serv	,474	,154	,493	3,078	,006

a. Dependent Variable: Salário_Y

Valores observados da estatística de teste (Teste T)

p-valores do teste T

13

Resultados do Ajuste do Modelo – *RLM Normal*:

Analyse / Regression / Linear

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	12,472	2,710		4,603	,000	6,853	18,091
	Idade	,304	,106	,459	2,868	,009	,084	,524
	Tempo_serv	,474	,154	,493	3,078	,006	,155	,793

a. Dependent Variable: Salário_Y

ICs de 95% para os parâmetros

14

Modelo de Regressão Linear Múltipla (RLM):

Análise dos resíduos e Avaliação das suposições básicas do modelo de RLM

Os métodos gráficos adotados no modelo de RLS para identificar violações das hipóteses também são válidos no caso do modelo de RLM.

<u>Análises Gráficas</u>:

- **(1) Gráfico da variável dependente *versus* cada uma das variáveis explicativas** → estudo da natureza e da força da relação entre as variáveis X_k`s e Y e detecção de valores discrepantes ou atípicos (*outliers*).

15

Modelo de Regressão Linear Múltipla (RLM):

Análise dos resíduos e Avaliação das suposições básicas do modelo de RLM - Análises gráficas (*continuação*):

- **(2) Gráfico de cada variável explicativa *versus* cada uma das outras variáveis explicativas** → identificação de colinearidade.
- **(3) Gráfico dos resíduos *versus* cada uma das variáveis explicativas** → avaliação da adequação do modelo de regressão em relação a cada variável explicativa, além de análise sobre possível variação na magnitude da variância dos erros do modelo no que se refere a cada variável explicativa.

16

Modelo de Regressão Linear Múltipla (RLM):

<u>Análises Gráficas</u>:

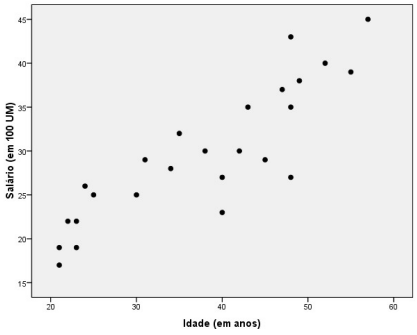
Análise dos resíduos e Avaliação das suposições básicas do modelo de RLM - Análises gráficas (*continuação*):

- **(4) Gráfico dos resíduos *versus* os valores ajustados** → análise da adequação do modelo de regressão, da hipótese de homocedasticidade e detecção de outliers.
- **(5) Histograma e/ou Gráfico dos Quantis (*QQ-Plot*) para os resíduos** → avaliação da hipótese de normalidade dos erros.

17

<u>Voltando ao exemplo de aplicação (modelo de RLM)</u>:

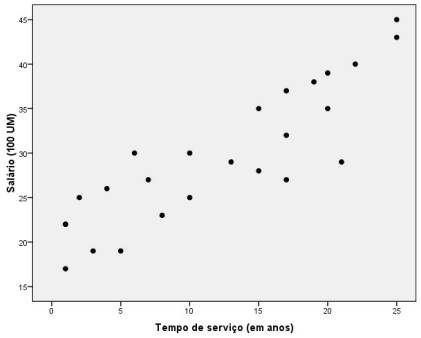
Figura 1.1: Gráfico de dispersão entre a idade e o salário dos funcionários.



18

Exemplo:

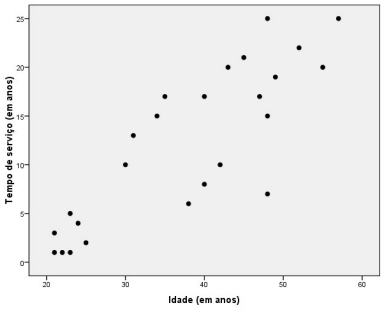
Figura 1.2: Gráfico de dispersão entre o tempo de serviço e o salário dos funcionários.



19

Exemplo:

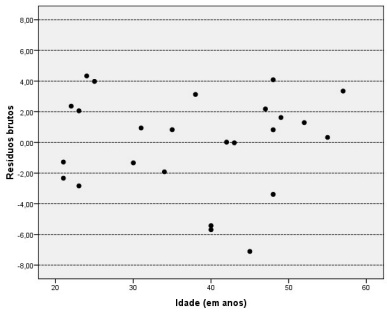
Figura 2: Gráfico de dispersão entre a idade e o tempo de serviço dos funcionários.



20

Exemplo:

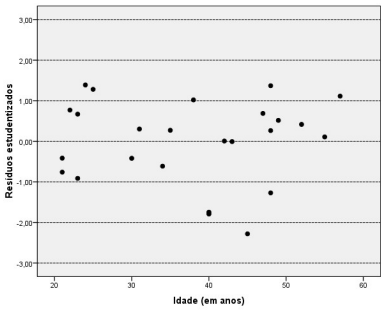
Figura 3.1: Gráfico de dispersão entre a idade e os resíduos brutos do modelo.



21

Exemplo:

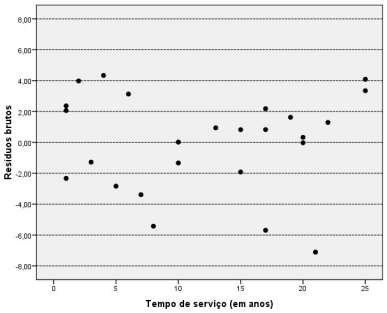
Figura 3.2: Gráfico de dispersão entre a idade e os resíduos estudentizados do modelo.



22

Exemplo:

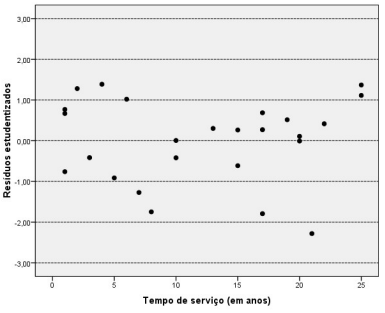
Figura 3.3: Gráfico de dispersão entre o tempo de serviço e os resíduos brutos do modelo.



23

Exemplo:

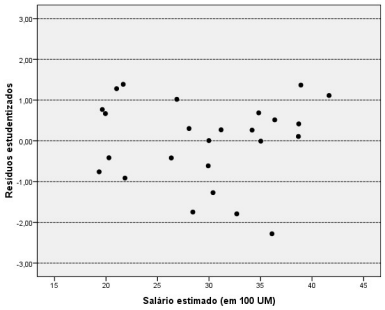
Figura 3.4: Gráfico de dispersão entre o tempo de serviço e os resíduos estudentizados do modelo.



24

Exemplo:

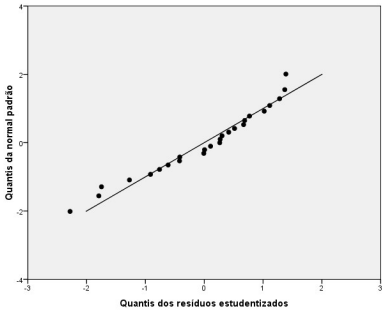
Figura 4: Gráfico de dispersão entre o salário estimado e os resíduos estudentizados do modelo.



25

Exemplo:

Figura 5: QQ-Plot (*normalidade*) para os resíduos estudentizados do modelo



26

Modelo de Regressão Linear Múltipla (RLM):
Análise dos resíduos e Avaliação das suposições básicas do modelo de RLM

Considerações finais:

- Os métodos descritos para contornar ou resolver as violações no modelo de RLS são também adotados no modelo de RLM.
- As transformações das variáveis podem ser feitas segundo os princípios discutidos no modelo de RLS para contornar eventuais violações no modelo de RLM.
  - ***Transformações da variável resposta*** → *Presença de heterocedasticidade (ou não normalidade).*
  - ***Transformações das variáveis explicativas*** → *Relações curvilíneas.*

27

Modelo de Regressão Linear Múltipla (RLM):
Análise dos resíduos e Avaliação das suposições básicas do modelo de RLM

Considerações finais (*continuação*):

- Quaisquer medidas usadas para contornar ou resolver possíveis violações devem ser examinadas através de gráficos de resíduos ou de outros métodos formais, de modo a avaliar a adequação do modelo para os dados transformados.

28

Aula prática – Exercício 1 ("*Saídas*"): Modelo de RLM com p-1=2 variáveis explicativas

A tabela a seguir fornece os salários semanais (em R$), a escolaridade e a horas semanais de trabalho de uma amostra de n=15 empregados de uma companhia.

O objetivo do estudo é avaliar a relação entre o salário semanal e as seguintes variáveis explicativas:

✓ *Anos de estudo*;

✓ *Hora semanal de trabalho*.

29

Dados sobre n=15 empregados de uma Companhia

Emp.	Anos de estudo	Hora semanal de trabalho	Salário semanal
1	4	10	350
2	8	14	400
3	12	16	470
4	10	26	550
5	15	31	620
6	7	12	380
7	6	13	290
8	10	21	490
9	11	26	580
10	13	24	610
11	12	23	560
12	8	12	420
13	11	19	450
14	12	19	510
15	5	11	380

30

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

a) Escreva a equação do modelo completo (contendo as variáveis X_1 e X_2) e descreva os seus termos e variáveis no contexto do problema.

b) Ajuste o modelo especificado na letra (a), interprete as somas dos quadrados da tabela ANOVA e o coeficiente de determinação. O que você pode concluir ? Justifique a sua resposta.

c) Avalie também para o modelo especificado na letra (a), a significância individual dos parâmetros dos modelo usando um teste estatístico de hipóteses apropriado a um nível de 5%. Você escolheria este modelo ? Justifique a sua resposta.

OBS: *Defina as 1) Hipóteses a serem testadas; 2) Estatística de teste; 3) Região Crítica; 4) Tomada de decisão.*

31

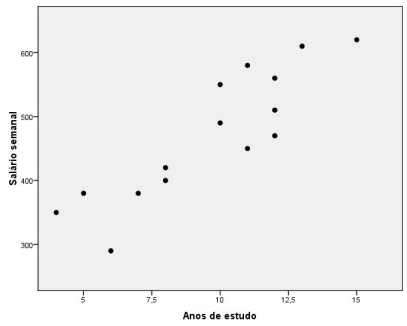
Aula prática - Exercício 1 ("*Saídas*"):

d) Para o modelo escolhido (selecionado), calcule uma medida global de qualidade do ajuste, interprete as estimativas dos parâmetros do modelo e avalie a sua significância estatística, considerando o nível de 5%. Além disso, avalie as hipóteses de *linearidade*, *normalidade*, *homocedasticidade* e *independência dos erros* usando a análise gráfica dos resíduos estudentizados.

32

Aula prática – Exercício 1:

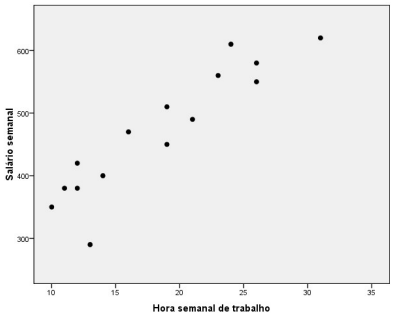
Figura 1: Gráfico de dispersão entre os anos de estudo e o salário semanal.



33

Aula prática – Exercício 1:

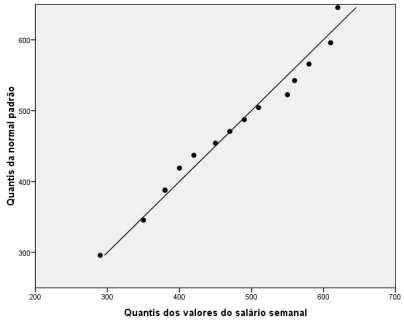
Figura 2: Gráfico de dispersão entre a hora semanal de trabalho e o salário semanal.



34

Aula prática – Exercício 1:

Figura 3: QQ-Plot (*normalidade*) para o salário semanal (Y) dos empregados da Companhia.



35

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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,941[a]	,885	,866	36,827

a. Predictors: (Constant), Hora_trabalho, Anos_estudo

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	125218,944	2	62609,472	46,165	,000[a]
	Residual	16274,389	12	1356,199		
	Total	141493,333	14			

a. Predictors: (Constant), Hora_trabalho, Anos_estudo

b. Dependent Variable: Salario_sem

36

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	184,884	31,762		5,821	,000	115,680	254,087
	Anos_estudo	11,746	5,835	,369	2,013	,067	-,968	24,460
	Hora_trabalho	9,369	2,825	,608	3,317	,006	3,215	15,524

a. Dependent Variable: Salario_sem

37

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	,920[a]	,846	,834	40,921

a. Predictors: (Constant), Hora_trabalho

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	119724,130	1	119724,130	71,496	,000[a]
	Residual	21769,203	13	1674,554		
	Total	141493,333	14			

a. Predictors: (Constant), Hora_trabalho

b. Dependent Variable: Salario_sem

38

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

Coefficients[a]

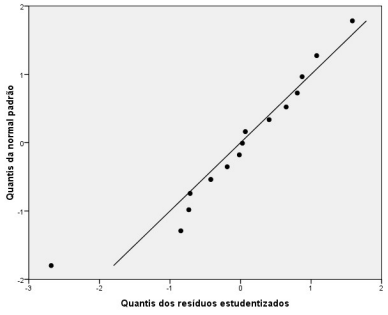
Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	208,876	32,714		6,385	,000	138,202	279,551
	Hora_trabalho	14,176	1,677	,920	8,456	,000	10,554	17,798

a. Dependent Variable: Salario_sem

39

Aula prática – Exercício 1:

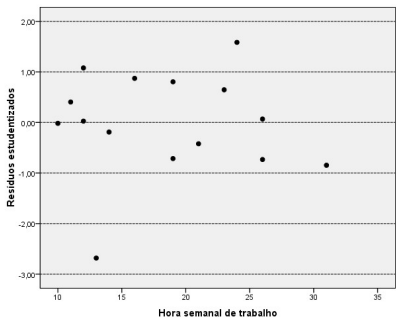
Figura 4: QQ-Plot (*normalidade*) para os resíduos estudentizados do modelo 2



40

Aula prática – Exercício 1:

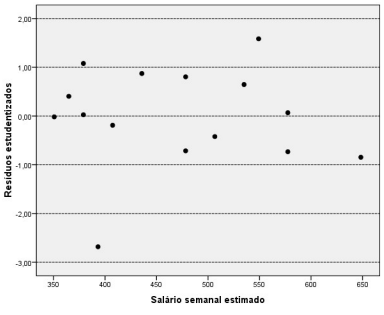
Figura 5: Gráfico de dispersão entre a hora de trabalho semanal e os resíduos estudentizados do modelo 2.



41

Aula prática – Exercício 1:

Figura 6: Gráfico de dispersão entre o salário semanal estimado e os resíduos estudentizados do modelo 2.



42

Aula prática – Exercício 2: ***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 deste 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.

Aula prática – Exercício 2 (*continuação*): ***Índice de distúrbio mental***

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:

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

Aula prática - Exercício 2 (*continuação*): ***Índice de distúrbio mental***

a) Represente usando gráficos apropriados a relação das variáveis consideradas no estudo. O que se pode concluir a partir desses gráficos ?

b) Escreva a equação do modelo completo (contendo as variáveis X_1 e X_2) e descreva os seus termos e variáveis no contexto do problema.

c) Ajuste o modelo definido na letra b), e use o teste de significância geral para responder a seguinte pergunta: Pelo menos uma das variáveis explicativas tem efeito estatisticamente significante ao nível de 5% ? **OBS:** *É preciso definir as Hipóteses a serem testadas, a Estatística de teste, Região Crítica e Tomada de decisão)*.

Aula prática - Exercício 2 (*continuação*): ***Índice de distúrbio mental***

d) Ainda para o modelo definido na letra b), interprete as estimativas dos parâmetros do modelo e avalie a sua significância individual usando um teste estatístico apropriado, considerando um nível de 5%. Em sua opinião, o sentido das relações encontradas é o esperado ? **OBS:** *É preciso definir as hipóteses a serem testadas, a Estatística de teste, Região Crítica e Tomada de decisão)*.

e) Calcule uma medida global de qualidade do ajuste do modelo final (interprete-a) e compare graficamente (e por meio de alguma medida apropriada) os índices de distúrbio mental observados e os estimados.

Aula prática - Exercício 2 (*continuação*): ***Índice de distúrbio mental***

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