Universidade Federal Fluminense (UFF)
Instituto de Matemática e Estatística (IME)
Departamento de Estatística (GET)
Disciplina: Modelos Lineares I
Professor: José Rodrigo de Moraes
2ª Lista de Exercícios – Data: 09/09/2019 (4ª feira)
Assunto: Inferência no modelo de RLS e análise dos resíduos.

1ª Questão: Mostre que os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ de mínimos quadrados (MQ) são não-viciados para os parâmetros β_0 e β_1, respectivamente.

2ª Questão: Encontre as variâncias dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ de MQ, e verifique de que parâmetro elas dependem.

3ª Questão: Considerando um modelo de RLS, demonstre que: $COV(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{X}\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ e diga qual o sentido da relação entre $\hat{\beta}_0$ e $\hat{\beta}_1$ para $\bar{X} > 0$, $\bar{X} < 0$ e $\bar{X} = 0$.

4ª Questão: Uma reação química é executada n vezes, e a temperatura (X), em ºC, e a produção (Y), em percentagem de um valor máximo teórico, são registradas para cada execução. As seguintes medidas estatísticas são registradas:

$\bar{X} = 65$ $\quad \bar{Y} = 29{,}05$ $\quad \sum_{i=1}^{12}(X_i - \bar{X})^2 = 6.032$ $\quad \sum_{i=1}^{12}(Y_i - \bar{Y})^2 = 835{,}42$

$\sum_{i=1}^{12}(X_i - \bar{X})(Y_i - \bar{Y}) = 1.988{,}4$

a) Calcule as estimativas de mínimos quadrados (MQ) dos parâmetros β_0 e β_1 do modelo, e escreva a equação do modelo ajustado.
b) Calcule a estimativa da variância do erro aleatório do modelo.
c) Determine os intervalos de confiança de 95% para os parâmetros β_0 e β_1.

d) Existe relação estatisticamente significante entre a temperatura e a produção, considerando um nível de significância de 5%.
e) Determine um intervalo de confiança de 95% para a produção média a uma temperatura de 40ºC.
f) Determine um intervalo de previsão para a produção de uma determinada reação a uma temperatura de 40ºC.

5ª Questão: Para verificar o efeito da variável X (horas de estudo) sobre a variável Y (nota da VS em Estatística Geral-II) foi realizado um estudo com 11 alunos que forneceu os seguintes pares (X_i, Y_i):
(3; 1,3), (7; 2,4), (5; 1,6), (2; 1,3), (9; 3,0), (7; 3,0), (3; 1,5), (5; 2,3), (8; 3,3), (2; 1,2), (1; 0,5)
Pede-se:

a) Construa um gráfico de dispersão entre X e Y e avalie se tem sentido ajustar um modelo de regressão linear normal aos dados. Nessa avaliação utilize também o coeficiente de correlação linear de Pearson.
b) Escreva de equação do modelo a ser ajustado descrevendo os seus termos/variáveis.
c) Ajuste um modelo de regressão linear normal e represente-o no gráfico de dispersão construído na letra (a).
d) Interprete as estimativas dos parâmetros do modelo (contexto do problema).
e) Verifique se existe associação estatisticamente significante entre as "horas de estudo" e o "desempenho na VS de EG-II". Defina as hipóteses a serem testadas, a estatística de teste (e o seu valor observado), a região crítica (RC) e a tomada de decisão, inclusive no contexto do problema.
f) Construa a tabela de Análise de Variância (ANOVA) e com base na estatística de teste F informe qual a conclusão a respeito da associação entre as horas de estudo (X) e o desempenho (nota) na disciplina (Y). A conclusão é a mesma da letra (e)?
g) Construa um IC de 95% para o efeito do tempo de estudo no desempenho do aluno na disciplina. Interprete-o. OBS: *Cheque os limites de confiança encontrados por meio do software R, aplicando as expressões conhecidas dos limites do IC para* β_1.

h) Construa um IC para o intercepto (β_0) do modelo. OBS*: Cheque os limites de confiança encontrados por meio do software R, aplicando as expressões conhecidas dos limites do IC para* β_0.

i) Calcule o coeficiente de determinação do modelo, e interprete-o no contexto do problema.

j) Usando os resíduos padronizados ou estudentizados, avalie a hipótese de normalidade dos erros do modelo. Use o histograma e o QQ-Plot.

k) Com base nos resultados do ajuste do modelo, você diria que o modelo se ajusta bem aos dados observados?

OBS: *Este exercício é para ser resolvido aplicando as fórmulas conhecidas dos estimadores de* β_0 *e* β_1*, das somas dos quadrados, das estatísticas de teste e IC; e também deve ser resolvido usando algum pacote estatístico, preferencialmente o programa R. Caso tenha acesso ao SPSS, pode ser utilizado para comparar os resultados obtidos com o R.*

6ª Questão: Os dados da tabela 1 se referem ao volume expiratório forçado (VEF) e a altura (com) de garotos na faixa etária de 10 a 14 anos. Estes dados foram coletados num hospital universitário por uma equipe multiprofissional.

Tabela 1: Dados sobre n=12 garotos.

Garoto	Altura	VEF
1	134	1,70
2	138	1,90
3	142	2,00
4	146	2,10
5	150	2,20
6	154	2,50
7	158	2,70
8	162	3,00
9	166	3,10
10	170	3,40
11	174	3,80
12	178	3,90

Pede-se:

a) Construa o gráfico de dispersão entre altura e VEF, e analise-o.

b) Calcule o coeficiente de correlação linear de Pearson entre altura e VEF. Avalie o sentido e o grau da relação entre essas duas variáveis.

c) Ajuste um modelo de RLS normal para explicar a variabilidade dos valores do VEF, a partir da altura dos garotos.

d) Verifique se existe associação estatisticamente significante entre a altura e VEF. Justifique a sua resposta.

e) Interprete o efeito estimado da altura sobre o VEF.

f) Avalie as hipóteses de *linearidade* e *normalidade* dos erros usando a análise gráfica dos resíduos padronizados ou estudentizados.

g) Escreva resumidamente as suas conclusões sobre a relação entre a altura e VEF, a partir dos resultados encontrados e das análises efetuadas. Se possível, escreva alguma recomendação no caso de existir violação da(s) hipótese(s) básica(s) do modelo.

7ª Questão: O dono de uma academia levantou o tempo de prática de natação (*em meses*) e o escore de satisfação com a capacidade cardiorespiratória (*numa escala de 0 a 100: quanto maior o escore, maior o nível de satisfação*) para um grupo de alunos que praticam natação no turno da noite (Tabela 2). O dono da academia quer avaliar o efeito o tempo de prática deste esporte sobre o nível de satisfação dos alunos com a sua capacidade cardiorespiratória.

a) Escreva a equação do modelo, descrevendo os seus termos e variáveis no contexto do problema.

b) Ajuste o modelo pelo método de mínimos quadrados (MQ) e interprete as estimativas dos parâmetros e o coeficiente de determinação do modelo (contexto).

c) Estime o nível médio de satisfação para o grupo de alunos que praticam natação a 1 ano e meio, e construa o seu respectivo IC de 95%. E para um grupo de alunos que praticam natação a 2 anos?

d) Estime o nível de satisfação de um aluno que pratica natação a 1 ano e meio, e construa o seu respectivo IC de 95%. E para um aluno que pratica natação a 2 anos?

e) Obtenha os intervalos de confiança e predição (*usando o software R*).

f) Construa um gráfico de dispersão para representar as bandas de confiança e predição, incluindo os dados observados e o modelo (reta) de regressão linear ajustado.

g) Avalie as hipóteses de *homocedasticidade*, *independência* e *normalidade* dos erros usando a análise gráfica dos resíduos padronizados ou estudentizados.

Tabela 2: Informações sobre n=20 alunos que praticam natação numa determinada academia.

Aluno	Tempo de natação (X)	Nível de satisfação (Y)
1	1,15	90
2	1,05	89
3	1,20	91
4	1,35	93
5	1,50	96
6	1,40	94
7	1,00	87
8	1,35	91
9	1,50	99
10	1,45	93
11	1,25	93
12	1,10	92
13	1,00	90
14	1,05	89
15	1,10	90
16	1,20	90
17	1,30	94
18	1,35	94
19	1,42	96
20	1,05	88
Total	24,77	1.839

$\sum_{i=1}^{n} X_i^2 = 31{,}2289$ $\sum_{i=1}^{n} Y_i^2 = 169.269$ $\sum_{i=1}^{n} X_i Y_i = 2.286{,}22$

8ª Questão: Em um estudo da relação entre o teor de oxigênio (medido em *partes de mil*) e a resistência máxima (medida em *ksi*), foram obtidos os seguintes resultados do ajuste de um modelo de RLS para n=29 soldas usando o pacote estatístico *SPSS*[1]:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,542[a]	,294	,268	5,84090

a. Predictors: (Constant), Teor_oxigenio

b. Dependent Variable: Resistência

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	383,094	1	383,094	11,229	,002[a]
	Residual	921,136	27	34,116		
	Total	1304,230	28			

a. Predictors: (Constant), Teor_oxigenio

b. Dependent Variable: Resistência

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	49,780	7,751		6,423	,000
	Teor_oxigenio	16,923	5,050	,542	3,351	,002

a. Dependent Variable: Resistência

[1] *SPSS - Statistical Package for the Social Sciences*

Figura 1: Gráfico de dispersão entre o teor de oxigênio (X) e os resíduos padronizados do modelo.

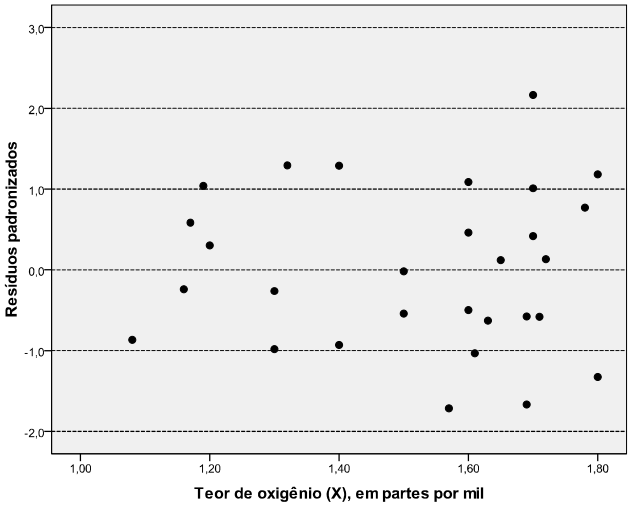


Figura 2: Gráfico de dispersão entre a resistência estimada (Ŷ) e os resíduos padronizados do modelo.

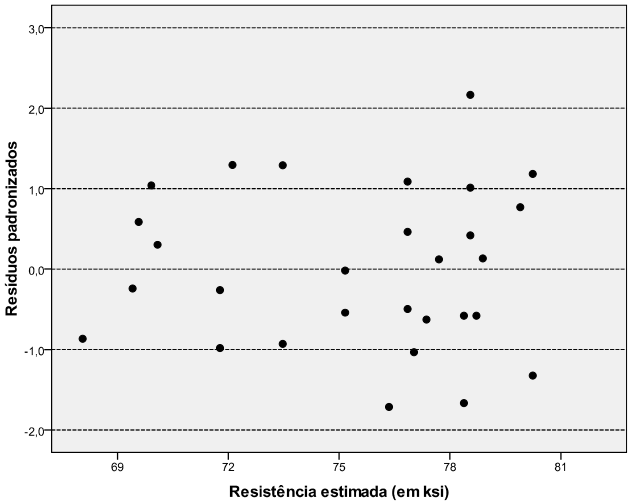
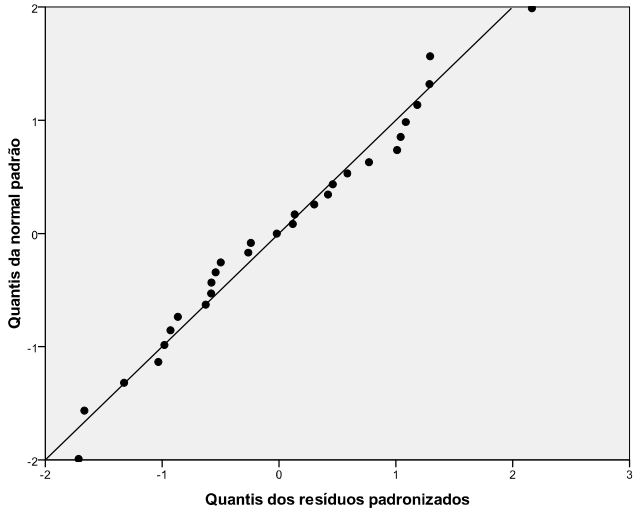


Figura 3: QQ Plot (*normalidade*) dos resíduos padronizados do modelo.



Usando as saídas fornecidas, pede-se:

a) Identifique a variável explicativa e a variável resposta do modelo, e escreva resumidamente qual o objetivo do estudo.

b) Escreva a equação do modelo teórico, descrevendo os seus termos e variáveis no contexto do problema. Inclua também na descrição as hipóteses básicas do modelo de RLS.

c) Escreva a equação do modelo ajustado pelo método de mínimos quadrados (MQ) e interprete as estimativas dos parâmetros do modelo (contexto).

d) Obtenha e interprete o coeficiente de correlação entre a resistência e o teor de oxigênio das soldas.

e) Obtenha e interprete o coeficiente de determinação do modelo (contexto).

f) Estime a variância do erro aleatório do modelo.

g) Determine e interprete os intervalos de confiança (IC) de 95% para os parâmetros do modelo.

h) Analise a tabela ANOVA, explicando o significado dos seus componentes no contexto do problema; e interprete o teste F.

i) Avalie a significância dos parâmetros do modelo. Você diria que existe relação estatisticamente significante entre o teor de oxigênio e a resistência das soldas ? Use o método do p-valor, mas é preciso justificar a sua resposta.

j) Determine o intervalo de confiança de 95% para a resistência média das soldas com teor de oxigênio de 1,7 partes por mil. **OBS:** Sabe-se que o teor médio de oxigênio é de 1,5197 partes de mil.

k) Determine o intervalo de predição de 95% para a resistência de uma solda cujo teor de oxigênio é 1,7 partes por mil.

l) Avalie as hipóteses de *homocedasticidade*, *independência* e *normalidade* dos erros usando a análise gráfica dos resíduos padronizados (Figuras 1, 2 e 3).

m) Em sua opinião, o modelo é adequado? Justifique a sua reposta.

n) Refaça o exercício considerando agora as saídas do **programa R**. Para tanto, os dados são fornecidos na tabela a seguir:

Teor de oxigênio (*partes de mil*)	Resistência (*ksi*)	Teor de oxigênio – *cont.* (*partes de mil*)	Resistência – *cont.* (*ksi*)
1,08	63,00	1,78	84,40
1,19	76,00	1,50	75,05
1,57	66,33	1,70	84,45
1,72	79,67	1,30	70,25
1,80	72,50	1,40	68,05
1,65	78,40	1,17	73,00
1,50	72,00	1,40	81,00
1,60	83,20	1,69	75,00
1,20	71,85	1,71	75,33
1,80	87,15	1,63	73,70
1,16	68,00	1,70	91,20
1,32	79,67	1,60	79,55
1,61	71,00	1,60	73,95
1,70	81,00	1,30	66,05
1,69	68,65		

9ª Questão: Os dados sobre o número de questões resolvidas de uma lista de exercícios indicada pelo professor da disciplina, contendo ao todo 50 exercícios, e a nota numa determinada prova final (0 a 100) são fornecidos na tabela 3.

a) Defina as variáveis X e Y, e classifique-as.

b) Construa o gráfico de dispersão entre X e Y, e analise-o.

c) Calcule o coeficiente de correlação linear de Pearson. Avalie o sentido e o grau da relação entre X e Y.

d) Ajuste um modelo de RLS normal para representar a relação entre X e Y, e avaliar a significância dessa relação usando um nível de significância de 5%.

e) Interprete as estimativas dos parâmetros do modelo, no contexto do problema.

f) Calcule uma medida de qualidade do ajuste do modelo da letra (d), e interprete-a no contexto do problema.

Tabela 3: Alunos matriculados numa determinada disciplina (n=28).

Aluno	Nº de questões resolvidas	Nota
1	1	0,6
2	1	1,6
3	1	0,5
4	1	1,2
5	2	2,0
6	2	1,3
7	2	2,5
8	3	2,2
9	3	2,4
10	3	1,2
11	4	3,5
12	4	4,1
13	4	5,1
14	5	5,7
15	6	3,4
16	6	9,7
17	6	8,6
18	7	4,0
19	7	5,5
20	7	10,5
21	8	17,5
22	8	13,4
23	8	4,5
24	9	30,4
25	11	12,4
26	12	13,4
27	12	26,2
28	12	7,4

10ª Questão: Usando as notas de estatística descritiva (*escala de 0 a 100*), o professor ajustou um modelo de RLS para explicar a nota da 2ª verificação, a partir da nota obtida na 1ª verificação da referida disciplina. Analise e interprete os resultados do ajuste do modelo ("*Saídas do SPSS*"), e faça a análise gráfica dos resíduos (Figuras 2, 3 e 4) à luz das hipóteses básicas do modelo. Elabore um pequeno relatório, com as principais conclusões obtidas por você.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3157,267	1	3157,267	39,759	,000[a]
	Residual	1985,252	25	79,410		
	Total	5142,519	26			

a. Predictors: (Constant), Nota_1

b. Dependent Variable: Nota_2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18,383	9,234		1,991	,058
	Nota_1	,774	,123	,784	6,305	,000

a. Dependent Variable: Nota_2

Figura 1: Gráfico de dispersão entre as notas da 1ª verificação (X) e 2ª verificação (Y), incluindo o modelo de RLS ajustado e o seu coeficiente de determinação.

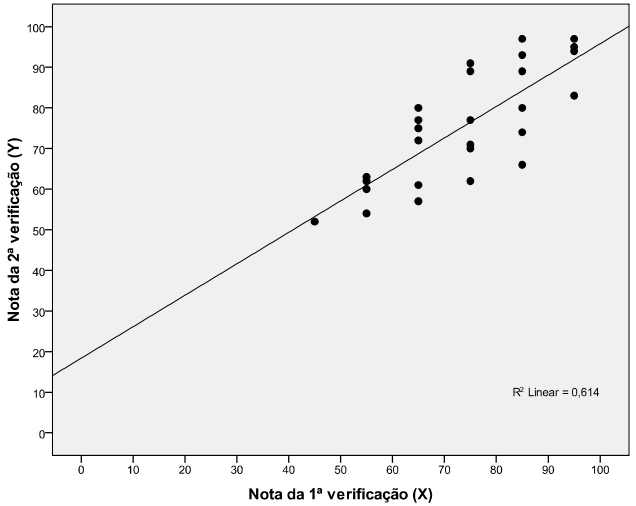


Figura 2: Gráfico de dispersão entre a nota da 1ª verificação (X) e os resíduos padronizados do modelo.

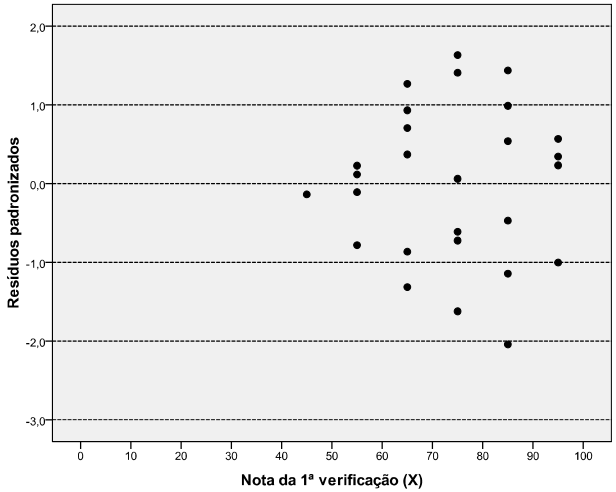


Figura 3: Gráfico de dispersão entre a nota estimada da 2ª verificação (Ŷ) e os resíduos padronizados do modelo.

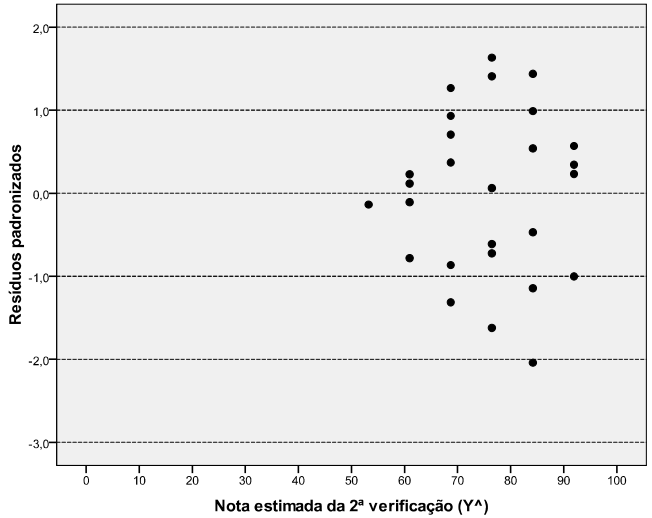
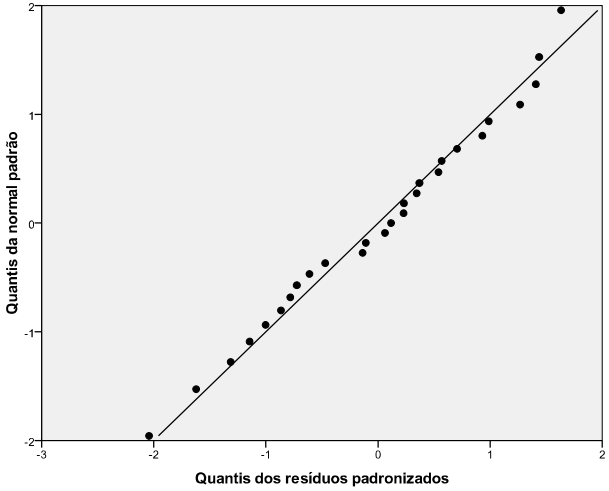


Figura 4: *QQ Plot* (normalidade) dos resíduos padronizados do modelo.



Respostas da 2ª Lista de Exercícios:

"Modelos Lineares I"

1ª Questão:

Tem que demonstrar que $E(\hat{\beta}_0) = \beta_0$ e $E(\hat{\beta}_1) = \beta_1$. *Consultar as notas de Aula do Prof. Dr. José Rodrigo, e/ou referências indicadas pelo professor.*

2ª Questão:

Consultar notas de Aula do Prof. Dr. José Rodrigo, e/ou referências indicadas pelo professor.

3ª Questão:

Consultar notas de Aula do Prof. Dr. José Rodrigo, e/ou referências indicadas pelo professor.

4ª Questão:

a) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 7{,}6233 + 0{,}32964\, X_i; \quad i = 1, 2, ..., 12$

b) 17,996

c) $IC_{\beta_0,95\%} = [-0{,}744; 15{,}991]$ e $IC_{\beta_1,95\%} = [0{,}208; 0{,}451]$

d) *O teste é significativo.*

e) $IC_{E(Y/X=45)\,;\,95\%} = [16{,}722\ ;\ 24{,}896]$

f) $IC_{Y/X=45\,;\,95\%} = [10{,}512\ ;\ 31{,}106]$

5ª Questão:

a) Sim, com base na observação do gráfico de dispersão, é possível dizer que tem sentido ajustar um modelo linear para os dados em questão. *Veja você mesmo !!!* Coeficiente de correlação linear de Pearson: R = 0,951

b) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i; \quad i = 1, 2, ..., 11$

c) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 0{,}452 + 0{,}316\, X_i; \quad i = 1, 2, ..., 11$

Escrever o significado de X_i, $\hat{Y}_i$, $\hat{\beta}_0$ e $\hat{\beta}_1$ no contexto.

d) $\hat{\beta}_0 = 0{,}452 \rightarrow$ é a nota estimada em EG-II no caso do aluno não ter estudado para a VS.

$\hat{\beta}_1 = 0{,}316 \rightarrow$ é o quanto aumenta a nota estimada em EG-II ao aumentar em o tempo de estudo em 1 hora.

e) *As hipóteses a serem testadas e a Região crítica (ℜC) ficam por conta do aluno.*

Valor observado da estatística de teste: t_{obs}=9,225. OBS: *É preciso ainda escrever a expressão da estatística de teste !!!*

Tomada de decisão: Como $t_{obs} \in$ ℜC rejeita-se H_0: β_1=0 ao nível de significância de 5%, ou seja, existe relação estatisticamente significante entre o "tempo de estudo (X)" e a "nota da VS (Y) de EG-II". Equivalentemente, usando o método do p-valor, temos que: p-valor < 0,001→ a relação é estatisticamente significante entre X e Y.

f) OBS: *É preciso construir a tabela ANOVA !!!*
SQM=7,404 (gl=1); SQRes= 0,786 (gl=9) e SQT= 8,187 (gl=10)
Valor observado da Estatística de Teste F: f_{obs}=85,108

Como o $f_{obs} \in$ RC (ou equivalentemente: p-valor<0,001<α=0,05), conclui-se, com base no teste F, que existe uma relação estatisticamente significante entre X (horas de estudo) e Y (nota da VS).

g) Pacote SPSS: $IC_{\beta_1,95\%}$ = [0,238 ; 0,393]

h) Pacote SPSS: $IC_{\beta_0,95\%}$ = [0,034 ; 0,870]

i) R^2=90,4%. OBS: *É preciso interpretar no contexto!!!*

j) *Por conta do aluno !!!*

k) *Por conta do aluno !!!*

6ª Questão:

a) Sugestão: Colocar altura no eixo X e VEF no eixo Y.

b) R= 0,998 → relação linear positiva forte. OBS: *É preciso justificar no contexto !!!*

c) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = -5{,}313 + 0{,}051 X_i$; $i = 1, 2, ..., 12$

d) Como p-valor < 0,001 → existe relação estatisticamente significante entre altura e FEV.

e) *Por conta do aluno !!! Ver notas de aula do Prof. Dr. José Rodrigo, e/ou referências indicadas*.

f) Há evidências de violação da hipótese básica de linearidade. Já a distribuição dos erros parece ser aproximadamente normal. **OBS:** *Cabe ressaltar que, para a análise de regressão ser válida, é necessário que as hipóteses básicas do modelo sejam pelo menos aproximadamente satisfeitas.*

7ª Questão:

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

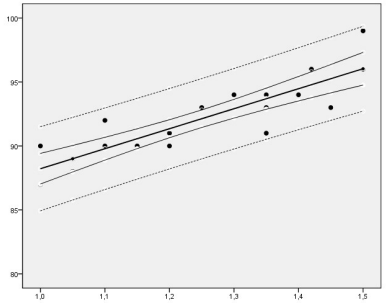
b) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 72{,}587 + 15{,}634 X_i$; $i = 1, 2, ..., 20$

c) $IC_{E(Y/X=1,5)\ ;\ 95\%} = [94{,}76\ ;\ 97{,}32]$

d) $IC_{Y/X=1,5\ ;\ 95\%} = [92{,}72\ ;\ 99{,}35]$

e) *Por conta do aluno !!! Usar o Programa R.*

f) **Figura:** Gráfico de dispersão entre o tempo de prática de natação (X) e o nível de satisfação (Y), incluindo o modelo de regressão ajustado e as bandas de confiança e predição.



g) Não se observou nenhuma violação das hipóteses básicas do modelo; sendo assim pode-se dizer que o modelo é adequado.

8ª Questão:

a) variável explicativa: *Teor de oxigênio* (X); variável resposta: *Resistência* (Y)

b) Modelo de RLS teórico:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$; $\varepsilon_i \sim N(0,\sigma^2) \forall\, i = 1,2,....n$ e $COV(\varepsilon_i, \varepsilon_j) = 0 \quad \forall\, i \neq j = 1,2,...,n$

c) Modelo de RLS ajustado: $\hat{Y}_i = 49{,}780 + 16{,}923\ X_i$; $i = 1, 2, ..., 29$

d) R=0,542 → moderada relação positiva entre X e Y.

e) R^2=0,294 → interpretação: *Ver notas de aula !!!*

f) $\hat{\sigma}^2 = \hat{VAR}(\varepsilon_i) = 34{,}116$. OBS: *Tem basicamente duas formas de obter esta estimativa !!!*

g) $IC_{\beta_0,95\%}$ = [33,877 ; 65,683] e $IC_{\beta_1,95\%}$ = [6,561 ; 27,285]

h) O teste F é significativo (p-valor = 0,002 < α=0,05) → existe relação significativa entre o teor de oxigênio e a resistência, ao nível de 5%.

i) Existe relação significativa entre o teor de oxigênio e a resistência.

j) $IC_{E(Y/X=1,7)\ ;\ 95\%} = [75{,}64\ ;\ 81{,}45]$ OBS: $t_{0{,}025;27} = 2{,}052$

k) $IC_{Y/X=1,7\ ;\ 95\%} = [66{,}22\ ;\ 90{,}88]$

l) Todas as hipóteses básicas do modelo são satisfeitas. OBS: *È necessário explicar !!!*

m) Sim, pois ...

n) Ajuste o modelo de RLS usando o programa R. Use os comandos: *lm*, *summary* e *anova.*

9ª Questão:

a) X → Nº de questões resolvidas e Y → nota

b) Sugestão: Usar a variável explicativa no eixo das abscissas e a variável resposta no eixo das ordenadas. OBS: *É preciso checar se a relação é do tipo linear ou não !!!*

c) R=0,736 → relação linear positiva forte. OBS: *É preciso justificar !!!*

d) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ = -1,452 + 1,558 X_i; $i = 1, 2, ..., 28$.

p-valor < 0,001 → existe relação estatisticamente significante entre X e Y.

e) *Por conta do aluno !!! Ver notas de aula do Prof. Dr. José Rodrigo e referências indicadas.*

f) R^2=54,2%. OBS: *É preciso interpretar essa medida no contexto do problema !!!*

10ª Questão: *Basta analisar as saídas do programa estatístico, mas é preciso representar o modelo, identificar os testes estatísticos incluindo todas as etapas até a tomada de decisão no contexto do problema.*