Universidade Federal Fluminense (UFF)
Instituto de Matemática e Estatística (IME)
Departamento de Estatística (GET)
Disciplina: Modelos Lineares I
Professor: Dr. José Rodrigo de Moraes
5ª Lista de Exercícios – Data: 13/11/2019 (4ª feira)
Assunto: Modelos de regressão linear com *dummy*.

1ª Questão: Os dados a seguir representam os pesos ao nascer (em gramas) de crianças nascidas de mães classificadas segundo seus hábitos em relação ao fumo.

Pede-se:

a) Proponha um modelo a ser ajustado aos dados observados e represente a sua equação descrevendo todos os termos e variáveis do modelo. Justifique a utilização do modelo.

b) Ajuste o modelo e verifique se existe efeito estatisticamente significante do fumo no peso das crianças. Interprete as estimativas dos parâmetros do modelo, e cheque a hipótese de normalidade dos erros usando o *QQ Plot* dos resíduos estudentizados.

c) Calcule uma medida global de qualidade do ajuste e interprete-a.

d) Identifique as observações discrepantes (outliers) e as mais influentes no ajuste do modelo, exclua-as e avalie o impacto deste procedimento nas estimativas dos parâmetros do modelo, na significância dos parâmetros e na qualidade global do ajuste. Sugestão: *Construa um gráfico entre a Distância de Cook e os resíduos estudentizados do modelo*.

e) Avalie a hipótese de normalidade dos erros para ambos os casos usando o *QQ Plot* e algum teste de normalidade que você conheça (*teste de Kolmogrov-Sminorv* ou *Shapiro-Wilk*, por exemplo).

Tabela 1: Peso ao nascer (em gramas) segundo os grupos de "fumantes" e "não fumantes".

Não fumantes	Fumantes
3.990	3.180
3.790	2.840
3.600	2.900
3.730	3.270
3.210	3.850
3.600	3.520
4.080	3.230
3.610	2.760
3.830	3.600
3.310	3.750
4.130	3.590
3.260	3.630
3.540	2.380
3.510	2.340
2.710	

2ª Questão: Os dados do banco a seguir contem informações sobre o peso do timo (em miligramas), sexo (0=fêmea, 1=macho), esterilização (0=não, 1=sim) e adrenalectomy (0=não, 1=sim) para 64 ratos da raça Wistar.

Pede-se:

a) Ajuste modelos de regressão com 1, 2 e 3 variáveis explicativas para explicar o peso do timo.

b) Utilizando o teste F de comparabilidade de modelos (Teste F parcial), avalie se o modelo com as três variáveis é mais adequado que os modelos com quaisquer duas variáveis explicativas.

c) Utilizando alguma medida de qualidade do ajuste apropriada (sem realização de teste), qual modelo você escolheria? Justifique a sua reposta.

d) Para o modelo escolhido (medida e teste), verifique se as hipóteses básicas do modelo foram satisfeitas, por meio da análise gráfica dos resíduos estudentizados.

Tabela 2: Informações sobre n=64 ratos da raça *Wistar*.

rato	sexo	esterilização	adrenalectomy	Peso do timo
1	1	0	0	26
2	1	0	0	24
3	1	0	0	25
4	1	0	0	31
5	1	0	0	20
6	1	0	0	19
7	1	0	0	27
8	1	0	0	27
9	0	1	1	52
10	1	1	1	45
11	0	1	0	32
12	0	1	1	52
13	1	1	0	44
14	1	1	0	41
15	1	1	1	48
16	0	1	1	48
17	1	0	1	31
18	0	1	0	45
19	0	0	1	38
20	1	1	1	39
21	1	1	1	34
22	0	0	1	47

Tabela 2 (*continuação*): Informações sobre 64 ratos da raça *Wistar*.

rato	sexo	esterilização	adrenalectomy	Peso do timo
23	0	0	0	35
24	1	1	0	43
25	0	0	1	44
26	1	0	1	39
27	0	0	0	29
28	0	0	0	36
29	1	1	0	39
30	0	0	1	42
31	0	1	0	33
32	0	1	0	34
33	1	0	1	21
34	0	0	0	50
35	0	1	1	45
36	0	1	0	39
37	1	0	1	26
38	1	0	1	35
39	0	1	0	34
40	1	0	1	25
41	0	0	0	31
42	0	0	1	39
43	0	1	1	40
44	0	1	0	42
45	0	0	0	35
46	1	1	1	41
47	1	1	1	38
48	0	0	1	48
49	1	1	0	36
50	1	0	1	32
51	1	1	0	31
52	0	0	1	44
53	0	1	1	46
54	1	0	1	44
55	0	1	1	56
56	0	0	1	35
57	1	1	1	55
58	1	1	1	44
59	0	1	1	55
60	0	0	0	48
61	1	1	0	32
62	1	1	0	41
63	0	0	0	48
64	0	1	0	35

3ª Questão: Nutricionistas e antropólogos da Nova Guiné realizaram um estudo para avaliar a associação entre a altura (em centímetros) e o conteúdo de proteína de crianças. Foram considerados dois grupos de participantes: um grupo formado por aquelas crianças submetidas a uma "dieta rica (1)" em proteína, enquanto o outro era composto por crianças submetidas a uma "dieta pobre (2)" em proteína.

a) Cheque a hipótese da normalidade da variável resposta usando o *QQ-Plot* ou algum teste de normalidade (*Teste Kolmogorov-Smirnov*). Qual a sua conclusão?

b) Especifique e ajuste um modelo estatístico para avaliar o efeito do conteúdo de proteína sobre a altura. OBS: *Considere o grupo 2 como 0.*

c) Especifique e ajuste um modelo estatístico para avaliar o efeito do conteúdo de proteína sobre a altura, levando em conta a idade das crianças.

d) Avalie a necessidade de se considerar a idade na associação de interesse entre o conteúdo de proteína e altura. Para tanto, utilize o teste F de comparabilidade de modelos. Caso seja importante a inclusão da idade no modelo, verifique se houve ou não mudança no efeito do tipo de proteína na altura da criança.

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

f) Identifique os *outliers*, exclua-os e ajuste um novo modelo. Com a exclusão dos outliers houve impacto no sentido e magnitude das associações? Verifique se os *outliers* identificados por você, também eram as observações mais influentes, em relação às demais observações do conjunto. Use a Distância de Cook.

Tabela 3: Altura e conteúdo de proteína de n=27 crianças

Criança (participante)	Grupo de dieta (proteína)	Idade	altura
1	2	1,0	61,0
2	2	2,8	76,0
3	2	1,8	69,0
4	2	3,0	74,0
5	2	2,4	72,0
6	2	1,0	63,4
7	2	1,3	65,0
8	2	2,0	67,9
9	2	2,0	68,5
10	2	3,0	77,0
11	2	1,5	66,0
12	2	0,7	55,0
13	2	0,4	52,0
14	2	0,2	51,0
15	1	1,0	66,0
16	1	1,0	69,0
17	1	1,8	82,0
18	1	0,5	54,3
19	1	0,2	54,0
20	1	2	80,3
21	1	2,5	93,2
22	1	3	94
23	1	2	83
24	1	1,4	73
25	1	2,7	94
26	1	2,5	91
27	1	0,8	63

4ª Questão: Considere uma amostra de 42 pacientes, onde a variável resposta é a lipoproteína de alta densidade (HDL). Existe a hipótese de que três variáveis são preditoras da HDL. São elas: 1) *colesterol total* (CT), 2) *triglicerídeo total* (TT) e 3) *sinking pré-beta* (SPB), a qual assume código 0 (ausente) ou código 1 (presente).

Pede-se:

a) Teste se as três variáveis isoladamente contribuem para a predição da HDL. Interprete os resultados do ajuste dos modelos (coeficientes do modelo, teste de significância dos parâmetros do modelo, medidas de qualidade do ajuste). Qual delas prediz melhor a HDL? Justifique a sua resposta.

b) Teste se as variáveis conjuntamente contribuem para a predição da HDL. Interprete os resultados do ajuste do modelo.

c) Supondo que o sinking pré-beta (SPB) é a variável de interesse no estudo da relação com a HDL. Verifique se CT ou TT ou ambas devem ser incluídas no modelo para melhorar a precisão ou a qualidade do ajuste do modelo.

d) Para o modelo selecionado, avalie as hipóteses de homocedasticidade, independência e normalidade dos erros foram atendidas. Use a análise gráfica dos resíduos estudentizados do modelo.

e) Identifique os *outliers*, exclua-os e ajuste um novo modelo. Com a exclusão dos outliers houve impacto no sentido e magnitude das associações?

Tabela 4: Amostra de n=42 pacientes.

Paciente	HDL	CT	TT	SPB
1	47	287	111	0
2	38	236	135	0
3	47	255	98	0
4	39	135	63	0
5	44	121	46	0
6	64	171	103	0
7	58	260	227	0
8	49	237	157	0
9	55	261	266	0
10	52	397	167	0
11	49	295	164	0
12	47	261	119	1
13	40	258	145	1
14	42	280	247	1
15	63	339	168	1
16	40	161	68	1
17	59	324	92	1
18	56	171	56	1
19	76	265	240	1
20	67	280	306	1
21	57	248	93	1
22	57	192	115	1
23	42	349	408	1
24	54	263	103	1
25	60	223	102	1
26	33	316	274	0
27	55	288	130	0
28	36	256	149	0
29	36	318	180	0
30	42	270	134	0
31	41	262	154	0
32	42	264	86	0
33	39	325	148	0
34	27	388	191	0
35	31	260	123	0
36	39	284	135	0
37	56	326	236	1
38	40	248	92	1
39	58	285	153	1
40	43	361	126	1
41	40	248	226	1
42	46	280	176	1

Respostas da 5ª Lista de Exercícios:

"Modelos Lineares I"

1ª Questão:

a) Modelo de regressão linear com 1 *dummy*:

$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \; ; \; \forall \; i = 1, 2, ..., n$

OBS: *É preciso definir os termos e variáveis do modelo.*

b) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} = 3.593{,}333 - 390{,}476 \; X_{i1} \; ; \; \forall i = 1, 2, ..., 29$

f_{obs} =5,868 (tabela ANOVA para o modelo completo); p-valor = 0,022 <0,05. Rejeita-se H_0 ao nível de significância de 5%, ou seja, o modelo reduzido não é tão adequado quanto ao modelo completo. Conclui-se, portanto, que existe relação estatisticamente significante entre o "fumo" e o "peso ao nascer (Y)" dos bebês.

c) R^2=17,9%

d) Observações discrepantes (*outliers*): obs 15 e 29

Observações mais influentes: obs 15, 29 e 28.

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} = 3.656{,}429 - 313{,}095 \; X_{i1} \; ; \; \forall i = 1, 2, ..., 26$

A significância dos parâmetros não mudou; R^2=19,7%

e) Para ambos os casos os resíduos aparentam ter distribuição aproximadamente normal.

2ª Questão:

a) <u>Modelos de regressão com 1 *dummy*</u>:

Modelo com "sexo": $\hat{Y}_i = 41{,}781 - 7{,}312 \; X_{i1}$

Modelo com "esterilização": $\hat{Y}_i = 34{,}406 + 7{,}437 \; X_{i2}$

Modelo com "adrenalectomty": $\hat{Y}_i = 34{,}750 + 6{,}750 \; X_{i3}$

OBS: *É preciso interpretar as estimativas dos parâmetros dos modelos !!!*

<u>Modelos de regressão com 2 *dummy*</u>:

Modelo com "sexo" e "esterilização": $\hat{Y}_i = 38{,}063 - 7{,}312 \; X_{i1} + 7{,}437 \; X_{i2}$

Modelo com "sexo" e "adrenalectomty": $\hat{Y}_i = 38{,}406 - 7{,}312 \; X_{i1} + 6{,}750 \; X_{i3}$

Modelo com "esterilização" e "adrenalectomty": $\hat{Y}_i = 31{,}031 + 7{,}437 \; X_{i2} + 6{,}750 \; X_{i3}$

OBS: *É preciso interpretar as estimativas dos parâmetros dos modelos !!!*

Modelo de regressão com 3 *dummy*:

$\hat{Y}_i = 34{,}688 - 7{,}312\ X_{i1} + 7{,}437\ X_{i2} + 6{,}750\ X_{i3}$

OBS: *É preciso interpretar as estimativas dos parâmetros no contexto do problema !!!*

b)

1º Teste – *Hipóteses a serem testadas:*

H_0: Modelo reduzido: modelo com "sexo" e "esterilização"

H_1: Modelo completo: modelo com "sexo", "esterilização" e "adrenalectomty".

Estatística de teste F:

$$f_{obs} = \frac{(SQ\mathrm{Re}s_0 - SQ\mathrm{Re}s_1)/(p-q)}{SQ\mathrm{Re}s_1/(n-p)} = \frac{(3.310{,}375 - 2.581{,}375)/(4-3)}{2.581{,}375/(64-4)} \cong 16{,}94$$

f_{obs} =16,94 ; $f_{1,60;\ 0,05}$ = 4,00. Rejeita-se H_0 ao nível de significância de 5%, ou seja, o modelo reduzido não é tão adequado quanto ao modelo completo.

2º Teste – *Hipóteses a serem testadas:*

H_0: Modelo reduzido: modelo com "sexo" e "adrenalectomty"

H_1: Modelo completo: modelo com "sexo", "esterilização" e "adrenalectomty".

Estatística de teste F:

f_{obs} =20,57 ; $f_{1,60;\ 0,05}$ = 4,00. Rejeita-se H_0 ao nível de significância de 5%, ou seja, o modelo reduzido não é tão adequado quanto ao modelo completo.

3º Teste – *Hipóteses a serem testadas:*

H_0: Modelo reduzido: modelo com "esterilização" e "adrenalectomty".

H_1: Modelo completo: modelo com ""sexo", "esterilização" e "adrenalectomty".

Estatística de teste F:

f_{obs} =19,89 ; $f_{1,60;\ 0,05}$ = 4,00. Rejeita-se H_0 ao nível de significância de 5%, ou seja, o modelo reduzido não é tão adequado quanto ao modelo completo.

Equação do modelo final:

$\hat{Y}_i = 34{,}688 - 7{,}312\ X_{i1} + 7{,}437\ X_{i2} + 6{,}750\ X_{i3}$

OBS: *É preciso descrever todos os termos/variáveis do modelo !!!*

c) Dos sete modelos de regressão ajustados, o modelo com 3 *dummy* também seria o escolhido (maior R^2_{ajust}=46,3%).

d) As hipóteses de homocedasticidade, independência e normalidade dos erros foram satisfeitas.

3ª Questão:

a) A hipótese de normalidade é aparentemente satisfeita (*QQ Plot*).

b)

Modelo com a variável *dummy* "grupo de proteína":

$\hat{Y}_i = 65{,}557 + 11{,}120\ X_{i1}\ ;\ \forall i = 1,2,...,27$

R^2=19,4%

c)

Modelo com a variável *dummy* "grupo de proteína" e "idade":

$\hat{Y}_i = 45{,}661 + 11{,}166\ X_{i1} + 12{,}058\ X_{i2};\ \forall i = 1,2,...,27$

R^2=90,7%

d)

Hipóteses a serem testadas:

H_0: Modelo reduzido: modelo com a *dummy* "grupo de proteína".

H_1: Modelo completo: modelo com a *dummy* "grupo de proteína" e "idade".

Estatística de teste F:

f_{obs} =183,699. Rejeita-se H_0 ao nível de significância de 5%, ou seja, o modelo reduzido não é tão adequado quanto ao modelo completo. *O que significa isso?*

e) *Por conta do aluno !!!*

f) *Outliers*: obs 4 e 18. As estimativas dos parâmetros do modelo não sofreram alteração substancial. Também são as observações mais influentes (relativamente às demais observações do conjunto).

4ª Questão:

a)

Modelo com "CT":

$\hat{Y}_i = 52{,}470 - 0{,}018 X_{i1}\ ;\ \forall\ i = 1,2,...,42$

Teste T-Student: t_{obs} =-0,636 (p-valor=0,528)

R^2=1,0%

Modelo com "TT":

$\hat{Y}_i = 46{,}245 + 0{,}010 X_{i2}\ ;\ \forall\ i = 1,2,...,42$

Teste T-Student: t_{obs} =0,431 (p-valor=0,669)

R^2=0,5%

Modelo com a dummy "SPB":

$\hat{Y}_i = 43{,}773 + 8{,}377 X_{i3}\ ;\ \forall\ i = 1,2,...,42$

Teste T-Student: t_{obs} =2,754 (p-valor=0,009)

R^2=15,9%

b)

Hipóteses a serem testadas:

H_0: Modelo reduzido: "modelo nulo"

H_1: Modelo completo: modelo com "CT", "TT" e "SPB"

Estatística de teste F (onde F ~ $F_{3,38}$*)*:

$$f_{obs} = \frac{(4.613{,}619 - 3.793{,}872)/(4-1)}{3.793{,}872/(42-4)} = \frac{273{,}249}{99{,}839} \cong 2{,}737$$ (ou p-valor=0,057)

Como p-valor é aproximadamente 0,05, pode-se rejeitar H_0 ao nível de 5%, e concluir que o modelo reduzido (modelo nulo) não é tão adequado quanto ao modelo completo. *Ou seja, pelo menos uma das variáveis explicativas contribui significativamente para a predição do HDL*.

c) As variáveis (CT, TT ou ambas) não melhoram substancialmente o poder de explicação do modelo. Tanto CT quanto TT devem ser excluídas do modelo por não estarem estatisticamente associadas com HDL, ao nível de 5% (Teste T). Para avaliar a significância do efeito de ambas as variáveis conjuntamente,

pode-se empregar o teste F de comparabilidade de modelos encaixados (a partir das somas dos quadrados dos resíduos): f_{obs}=0,423 (F~$F_{2,38}$), ou p-valor=0,658. *Facilmente obtido no excel:* $DISTF(f_{obs};p\text{-}q;n\text{-}p)$.

d) Equação do modelo escolhido:

$\hat{Y}_i = 43{,}773 + 8{,}377 X_{i3}\ ;\ \forall\ i = 1,2,...,42$

Teste T-Student: t_{obs} =2,754 (p-valor=0,009) R^2=15,9% e R^2ajust=13,8%

As hipóteses básicas do modelo aparentemente foram atendidas.

e) Equação do modelo ajustado (excluindo os outliers 6 e 19):

$\hat{Y}_i = 42{,}810 + 8{,}085 X_{i3}\ ;\ \forall\ i = 1,2,...,40$

Teste T-Student: t_{obs} =2,948 (p-valor=0,005) R^2=18,6% e R^2ajust=16,5%

As estimativas não sofreram alteração substancial (redução menor do que 10%) e as hipóteses básicas do modelo atendidas.