Universidade Federal Fluminense (UFF)

Instituto de Matemática e Estatística (IME)

Departamento de Estatística (GET)

Disciplina: Modelos Lineares I

Professor: José Rodrigo de Moraes

4ª Lista de Exercícios – Data: 04/11/2019 (2ª feira)

Assunto: Problemas na análise de regressão linear: *Multicolinearidade, heterocedasticidade, outliers e observações influentes*.

1ª Questão: Os dados sobre poluição ambiental em 41 cidades americanas se são apresentados na tabela 1. A variável dependente é a concentração média anual de dióxido sulfúrico (SO_2), expressa em microgramas por metro cúbico. As variáveis explicativas são seis: 1) *Temp*: temperatura anual média em ºF; 2) *Manuf*: número de estabelecimentos de manufatura que empregam 20 ou mais trabalhadores; 3) *Pop*: tamanho da população em miles, pelo censo de 1970; 4) *Vento*: média anual da velocidade do vento em milhas por hora; 5) *Precip*: precipitação anual média em polegadas; 6) *N_dias*: número médio de dias com precipitação por ano. O objetivo do estudo é identificar as variáveis que estão relacionadas com a concentração de dióxido sulfúrico (SO_2) usando alguma análise estatística apropriada.

a) Diga que tipo de análise estatística poderia ser realizada para atingir o objetivo do estudo. Justifique a sua resposta.

b) Obtenha os gráficos de dispersão e os coeficientes de correlação linear de Pearson entre SO_2 com cada uma das variáveis independentes. Analise-os.

c) Obtenha a matriz de correlações entre os possíveis pares de variáveis explicativas. Comente.

d) Ajuste o modelo de RLM para explicar a variabilidade de SO2, a partir das seis variáveis explicativas consideradas no estudo. Analise os testes de significância individual e geral dos parâmetros do modelo. Você diria que há a indícios de multicolinearidade entre as variáveis? Justifique a sua resposta.

e) Para confirmar a sua resposta, obtenha o "fator de inflação da variância" (VIF) para cada uma das variáveis explicativas do modelo ajustado na letra (d). Qual a sua conclusão final?

f) Em caso de existência de multicolinearidade, tome alguma medida para contornar este problema. Explique o uso da medida escolhida.

g) Escreva a equação do modelo que você selecionou e interprete as estimativas dos parâmetros do modelo. Calcule ainda alguma medida de qualidade do ajuste, e interprete-a no contexto do problema.

Tabela 1: Dados sobre 41 cidades americanas

Cidade	SO2	Temp	Manuf	Pop	Vento	Precip	N_dias
1	10	70,3	213	582	6,0	7,05	36
2	13	61,0	91	132	8,2	48,52	100
3	12	56,7	453	716	8,7	20,66	67
4	17	51,9	454	515	9,0	12,95	86
5	56	49,1	412	158	9,0	43,37	127
6	36	54,0	80	80	9,0	40,25	114
7	29	57,3	434	757	9,3	38,89	111
8	14	68,4	136	529	8,8	54,47	116
9	10	75,5	207	335	9,0	59,80	128
10	24	61,5	368	497	9,1	48,34	115
11	110	50,6	3344	3369	10,4	34,44	122
12	28	52,3	361	746	9,7	38,74	121
13	17	49,0	104	201	11,2	30,85	103
14	8	56,6	125	277	12,7	30,58	82
15	30	55,6	291	593	8,3	43,11	123
16	9	68,3	204	361	8,4	56,77	113
17	47	55,0	625	905	9,6	41,31	111
18	35	49,9	1064	1513	10,1	30,96	129
19	29	43,5	699	744	10,6	25,94	137
20	14	54,5	381	507	10,0	37,00	99
21	56	55,9	775	622	9,5	35,89	105
22	14	51,5	181	347	10,9	30,18	98
23	11	56,8	46	244	8,9	7,77	58
24	46	47,6	44	116	8,8	33,36	135
25	11	47,1	391	463	12,4	36,11	166
26	23	54,0	462	453	7,1	39,04	132
27	65	49,7	1007	751	10,9	34,99	155
28	26	51,5	266	540	8,6	37,01	134
29	69	54,6	1692	1950	9,6	39,93	115
30	61	50,4	347	520	9,4	36,22	147
31	94	50,0	343	179	10,6	42,75	125
32	10	61,6	337	624	9,2	49,10	105
33	18	59,4	275	448	7,9	46,00	119
34	9	66,2	641	844	10,9	35,94	78
35	10	68,9	721	1233	10,8	48,19	103
36	28	51,0	137	176	8,7	15,17	89
37	31	59,3	96	308	10,6	44,68	116
38	26	57,8	197	299	7,6	42,59	115
39	29	51,1	379	531	9,4	38,79	164
40	31	55,2	35	71	6,5	40,75	148
41	16	45,7	569	717	11,8	29,07	123

2ª Questão: Os dados sobre concentração de resíduos de PCB (*Polychlorinated biphenil*), em ppm, encontrados em peixes de um lago e a idade dos peixes são fornecidos na tabela 2.

a) Faça o gráfico de dispersão e calcule o coeficiente de correlação linear de Pearson entre a concentração de resíduos de PCB e idade dos peixes; e avalie indícios de violação de alguma das hipóteses básicas do modelo.

b) Ajuste um modelo de regressão linear normal para explicar a variabilidade da concentração de resíduos de PCB, a partir da idade dos peixes. Faça o gráfico de dispersão entre a idade e os resíduos estudentizados do modelo. Aproveite a oportunidade e construa também o *QQ Plot* destes resíduos. Qual a sua conclusão com base nestes dois gráficos?

c) Aplique o teste de White para avaliar se o modelo ajustado na letra (b) é heterocedástico. Use α=5%

d) Aplique o logaritmo (neperiano) da variável PCB e ajuste um novo modelo de regressão linear normal. Com base nos gráficos de dispersão (idade *versus* lnPCB; e idade *versus* resíduos estudentizados) e no teste de White (α=5%), o modelo continua sendo heterocedástico? E quanto à normalidade dos erros deste modelo, o que você tem a dizer?

e) Houve aumento do poder explicativo do modelo com lnPCB, relativamente ao modelo com PCB? Em caso afirmativo, qual foi o acréscimo?

f) Você conhece outra forma de corrigir o problema de heterocedasticidade?

Tabela 2: Dados sobre 28 peixes de um lago

Peixe	Idade	PCB	Peixe	Idade	PCB
1	1	1,0	**15**	6	3,5
2	1	2,0	**16**	6	9,8
3	1	1,0	**17**	6	8,5
4	1	1,0	**18**	7	4,5
5	2	2,0	**19**	7	6,0
6	2	1,5	**20**	7	11,0
7	2	3,0	**21**	8	16,0
8	3	2,5	**22**	8	12,0
9	3	2,2	**23**	8	4,0
10	3	1,5	**24**	9	25,0
11	4	3,0	**25**	11	12,0
12	4	4,0	**26**	12	13,0
13	4	5,0	**27**	12	26,0
14	5	5,8	**28**	12	7,5

3ª Questão: Com o objetivo de investigar a relação entre as notas de estatística descritiva (numa escala de 0 a 100), o professor registrou no início do semestre as notas da 1ª verificação dos seus 27 alunos. No fim do semestre anotou as notas da 2ª verificação como pode ser visto na tabela abaixo:

Tabela 3: Notas das 1ª e 2ª verificação para 27 alunos matriculados.

1ª Nota	45	55	55	55	55	65	65	65	65	65	65	75	75	75	75	75	75	85	85	85	85	85	85	95	95	95	95
2ª Nota	52	54	63	60	62	57	72	77	80	61	75	62	77	91	71	89	70	89	93	97	74	80	66	94	97	83	95

a) Ajuste o modelo de regressão linear e avalie a existência de heterocedasticidade por meio do gráfico de dispersão entre a nota da 1ª verificação e os resíduos estudentizados do modelo.

b) Utilizando o teste de White considerando o nível de significância de 5%, qual a sua opinião? É necessário definir todas as etapas do teste: 1ª) *Hipóteses a serem testadas*, 2ª) *Estatística de teste*, 3ª) *Região Crítica*, 4ª) *Tomada de decisão*).

3)Interprete as estimativas dos parâmetros do modelo (*sem heterocedasticidade*) e o coeficiente de determinação do modelo.

4ª Questão: Considerando as quantidades *per capita* de telefones (X) e de arrecadação do imposto de circulação de mercadoria (Y), em 13 localidades de um estado brasileiro.

Tabela 4: Dados referentes a n=13 localidades de um estado brasileiro

Localidade	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Qde de tels	42	44	48	53	56	58	58	65	68	70	77	86	138
ICM	1,95	2,39	2,5	3,22	3,63	3,54	3,65	4,49	5,78	5,4	1,14	13,94	12,66

a) Faça o gráfico de dispersão entre X e Y. Faça comentários com relação a presença de oultiers.

b) Ajuste o modelo de regressão linear (modelo 1) para os dados da tabela 4 e obtenha os resíduos estudentizados do modelo e os valores da distância de Cook (D). Analisando os gráficos de dispersão identifique os outliers e as observações mais influentes no conjunto de dados. Avalie ainda se a hipótese de normalidade dos dados é satisfeita.

c) Supondo que os outliers identificados por você são resultantes de erros de mensuração ou informação, ou de fato casos excepcionais que dificilmente ocorreriam em outra localidade da população, exclua-os e ajuste o modelo (modelo 2). Repita as análises gráficas dos resíduos estudentizados e dos valores da distância de Cook para o modelo 2. A localidade de número 13 continua sendo uma observação com alta influência? Quais as conseqüências de se manter essa localidade na análise e selecionar o modelo 2?

d) Por fim, elimine a localidade de número 13, e ajuste um novo modelo (modelo 3) para os dados restantes. Verifique se as hipóteses de homocedasticidade e normalidade estão satisfeitas? Qual a sua conclusão?

5ª Questão: Considerando os dados da tabela 5, resolva os itens solicitados abaixo:

Tabela 5: Amostra aleatória simples de tamanho n=19 indivíduos.

Id	Y	X_1	X_2	X_3
1	42	162	23	3
2	34	162	23	8
3	28	162	30	5
4	22	162	30	8
5	20	172	25	5
6	15	172	25	8
7	12	172	30	5
8	4,3	172	30	8
9	19	167	28	7
10	6,4	177	28	7
11	38	157	28	7
12	18	167	33	7
13	26	167	23	7
14	9,9	167	28	10
15	25	167	28	4
16	14	177	20	7
17	15	177	20	7
18	16	160	34	8
19	20	160	34	8

a) Faça o gráfico de dispersão entre Y e cada uma das variáveis explicativas X_1, X_2 e X_3, e avalie o sentido e a força da relação.

b) Ajuste o modelo de RLM e avalie a significância dos parâmetros do modelo (α=5%), e calcule e interprete o coeficiente de determinação do modelo.

Obtenha ainda os fatores de inflação de variância (VIF) cada variável explicativa e avalie a existência de multicolineradade.

c) Faça o gráfico de dispersão dos valores ajustados *versus* resíduos estudentizados do modelo, e identifique os outliers.

d) Faça o gráfico de dispersão das medidas de alavancagem (*"leverage"*) *versus* a ordem das observações; e identifique as observações com alta alavancagem.

e) Faça o gráfico de dispersão das distâncias de Cook (D) *versus* a ordem das observações; e identifique as observações com alta influência no ajuste do modelo de regressão. Se desejar, construa também o *box-plot*, para identificar aquelas observações que possuem valores de D que são superiores aos valores das demais observações.

f) Faça o gráfico de dispersão das distâncias de Cook (eixo das abscissas) *versus* os resíduos estudentizados (eixos das ordenadas), e cheque se todas as observações influentes são consideradas outliers.

g) Se você excluísse estas observações com alta influência na letra (e), que tipo de mudança ocorreria nas estimativas dos parâmetros do modelo e na qualidade global do ajuste?

6ª Questão: Para cem comunidades carentes de um estado, registrou o *percentual (%) domicílios inadequados quanto à qualidade da moradia* e o *percentual (%) de pessoas satisfeitas com local onde mora*. Os dados se encontram na Tabela 6.

a) Ajuste o modelo de RLS e avalie a significância dos parâmetros do modelo (α=5%), e avalie se as hipóteses de homocedasticidade, independência e normalidade dos erros são satisfeitas usando os resíduos estudentizados do modelo. Analise cada gráfico que você construiu.

b) Faça o gráfico de dispersão das distância de Cook *versus* resíduos estudentizados do modelo. Identifique os outliers e as observações mais influentes no conjunto de dados.

c) Se você excluir as duas observações mais influentes do conjunto de dados, há alteração substancial das estimativas dos parâmetros do modelo e da qualidade global do modelo? Reavalie as hipóteses básicas do modelo.

d) Qual modelo você escolheria? Justifique a sua resposta.

Tabela 6: Dados referentes a cem comunidades de baixa renda

Comunidade	% domicílios inadequados	% pessoas satisfeitas	Comunidade	% domicílios inadequados	% pessoas satisfeitas
1	12,0	34,9	**51**	12,0	35,0
2	12,6	34,8	**52**	10,6	35,1
3	12,1	34,9	**53**	10,2	35,0
4	11,5	34,9	**54**	12,4	34,6
5	15,1	34,5	**55**	10,8	34,5
6	13,4	34,8	**56**	12,8	34,6
7	12,9	34,9	**57**	12,4	34,9
8	10,9	35,0	**58**	10,3	34,8
9	12,9	34,7	**59**	13,3	34,5
10	12,4	34,8	**60**	11,3	35,2
11	13,7	34,9	**61**	11,9	34,4
12	11,8	35,0	**62**	11,4	34,7
13	13,0	34,5	**63**	10,1	35,2
14	10,8	34,9	**64**	12,0	34,6
15	11,3	34,9	**65**	12,1	34,6
16	12,4	34,7	**66**	12,0	35,0
17	9,2	35,3	**67**	12,6	34,6
18	13,7	34,5	**68**	12,1	34,5
19	11,4	35,1	**69**	12,6	34,8
20	10,4	34,8	**70**	13,3	35,0
21	11,2	35,0	**71**	13,3	34,9
22	11,3	34,8	**72**	12,5	34,7
23	12,0	34,6	**73**	12,1	34,9
24	13,8	34,4	**74**	12,1	34,7
25	11,1	34,7	**75**	12,5	34,5
26	11,7	35,0	**76**	12,6	34,6
27	11,0	34,6	**77**	12,2	35,0
28	10,9	34,9	**78**	11,1	35,3
29	11,0	35,1	**79**	10,5	34,9
30	11,8	34,6	**80**	12,9	34,2
31	10,7	35,0	**81**	11,0	35,1
32	11,6	34,8	**82**	10,6	35,0
33	13,0	34,7	**83**	12,7	34,7
34	11,7	34,8	**84**	10,4	34,9
35	11,4	34,7	**85**	13,2	34,7
36	12,2	34,7	**86**	12,3	34,9
37	11,1	35,0	**87**	11,4	34,9
38	12,6	34,6	**88**	13,6	34,6
39	11,8	34,7	**89**	12,7	34,5
40	10,8	34,8	**90**	14,0	34,9
41	11,7	34,9	**91**	11,9	35,1
42	10,7	34,6	**92**	11,6	35,0
43	13,4	34,6	**93**	13,2	34,6
44	11,3	34,7	**94**	13,0	34,6
45	9,8	34,9	**95**	11,7	34,7
46	11,5	34,6	**96**	12,1	34,7
47	13,6	34,3	**97**	10,8	35,4
48	13,1	34,6	**98**	10,6	35,0
49	11,4	34,8	**99**	11,5	34,9
50	10,7	34,6	**100**	13,9	34,5

7ª Questão: A renda mensal média de vendas de refeições (Y) e os gastos mensais com propaganda (X) foram registrados por 30 gestores de restaurantes. Responda as questões abaixo com objetivo de estabelecer a relação entre a renda de vendas e os gastos mensais com propagandas.

a) Procure indícios de heterocedasticidade visualizando o gráfico de dispersão entre X e Y. O que você acha?

b) Ajuste o modelo de regressão linear e construa o gráfico de dispersão dos resíduos estudentizados *versus* valores ajustados e X. E agora qual a sua opinião?

c) Utilizando o teste de White considerando o nível de significância de 5%, qual a sua conclusão final? É necessário definir todas as etapas do teste: 1ª) *Hipóteses a serem testadas*, 2ª) *Estatística de teste*, 3ª) *Região Crítica*, 4ª) *Tomada de decisão*).

d) Em caso de existência, corrija o problema de heterocedasticidade e ajuste um novo modelo.

Tabela 7: Dados sobre n=30 agências turísticas

Restaurante	Vendas	Gastos com propaganda	Restaurante	Vendas	Gastos com propaganda
1	81	3	**16**	147	13
2	73	3	**17**	179	15
3	72	3	**18**	166	15
4	91	5	**19**	181	15
5	99	5	**20**	178	15
6	127	9	**21**	185	15
7	114	9	**22**	156	15
8	116	9	**23**	173	16
9	123	9	**24**	189	16
10	131	9	**25**	192	17
11	141	11	**26**	203	19
12	151	12	**27**	192	19
13	147	12	**28**	219	19
14	131	12	**29**	214	19
15	145	12	**30**	185	19

Respostas da 4ª Lista de Exercícios:

"Modelos Lineares I"

1ª Questão:

a) Análise de regressão linear múltipla (RLM).

b) *Por conta do aluno !!!*

c) Matriz de correlações entre pares de variáveis explicativas:

	Temp	Manuf	Pop	Vento	Precip	N_dias
Temp	1	-0,19	-0,06	-0,35	0,39	-0,43
Manuf		1	0,96	0,24	-0,03	0,13
Pop			1	0,21	-0,03	0,04
Vento				1	-0,01	0,16
Precip					1	0,50
N_dias						1

d) Ficou difícil avaliar a presença de (multi)colinearidade neste caso, sobretudo devido ao fenômeno em questão. Cabe ressaltar que apesar das variáveis "*Manuf*" e "*Pop*" serem fortemente correlacionadas (análise bivariada), na análise multivariada as variáveis apresentaram efeito significativo (p-valor<5%). Recomendação: Calcular os fatores de inflação de variância (VIF) para todas as variáveis do modelo.

e) "*Manuf*" e "*Pop*" apresentam VIF> 10 → existe colinearidade entre estas variáveis.

f) Excluir a variável "*Pop*" e reajustar o modelo com as demais variáveis.

g) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} = 77{,}237 - 1{,}048\, X_{i1} + 0{,}024\, X_{i2}$; $\forall \quad i = 1, 2, ..., 41$

Todos os parâmetros são significativos (p-valor < 5%).

R^2=51,6%

2ª Questão:

a) R=0,764 (relação positiva). Pelo gráfico de dispersão entre X e Y, parece haver indícios de violação da hipótese de homocedasticidade, mas cabe maior investigação.

b) Modelo com "PCB": $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} = -1{,}068 + 1{,}446\, X_{i1}$; $\forall \quad i = 1, 2, ..., 28$

Intercepto: p-valor=0,501 > 0,05

Idade: p-valor <0,001 < 0,05

Coeficiente de determinação do modelo: R^2=58,4%

Gráfico de dispersão entre a idade e os resíduos estudentizados do modelo → violação da hipótese de homocedasticidade.

QQ Plot dos resíduos estudentizados → violação da hipótese de normalidade.

c) Teste de White: $w_{obs} = nR^2 = 28 \cdot 0{,}326 = 9{,}128 > \chi^2_{0{,}05;2} = 5{,}991 \rightarrow$ o modelo é heterocedástico.

d) Modelo com "lnPCB": $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} = 0{,}220 + 0{,}234\, X_{i1}$; $\forall \quad i = 1, 2, ..., 28$

Intercepto: p-valor=0,214 > 0,05

Idade: p-valor <0,001 < 0,05

Coeficiente de determinação do modelo: R^2=75,0%

Gráficos de dispersão (idade *versus* lnPCB; e idade *versus* resíduos estudentizados) → ambos os gráficos fornecem índicos que a hipótese de homocedasticidade é satisfeita.

Teste de White: $w_{obs} = nR^2 = 28 \cdot 0{,}202 = 5{,}656 < \chi^2_{0{,}05;2} = 5{,}991 \rightarrow$ o modelo é homocedástico.

QQ Plot dos resíduos estudentizados → a hipótese de normalidade é satisfeita.

e) Verificou também aumento do poder explicativo, e este aumento foi de 28,4%.

f) *Por conta do aluno !!! Mostre o procedimento de correção.*

3ª Questão:

a) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1\, X_i = 18{,}383 + 0{,}774\, X_i$; $\quad i = 1, 2, ..., 27$

Gráfico de dispersão entre X e os resíduos estudentizados → parece sugerir algum (leve) aumento proporcional na variabilidade dos resíduos em função idade (X).

b) Teste de White: $w_{obs} = nR^2 = 27 \cdot 0{,}204 = 5{,}508 < \chi^2_{0{,}05;2} = 5{,}991 \rightarrow$ não existe problema de heterocedasticidade

4ª Questão:

a) As observações 11, 12 e 13 se encontram afastadas da grande massa de dados. Este gráfico sugere que são outliers em potencial, mas é possível confirmar na análise dos resíduos estudentizados [-2,+2].

b)

Modelo 1: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = -3{,}252 + 0{,}123 X_i ; \quad i = 1, 2, ..., 13$

R^2=61,6% e R^2ajust=58,1%

Intercepto: p-valor=0,146 < 0,05

Nº de tels: p-valor=0,001 < 0,05

Análise gráfica dos resíduos estudentizados → observações 11 e 12 são *outliers*.

Análise gráfica dos valores da distância de Cook (D) → a observação 13 é uma observação altamente influente (D_{13}=1,298>1), pois o valor D_{13} é maior do que 1, e bem superior aos valores de D das demais observações (inclusive das observações também influentes 11 e 12). Isto pode ser melhor visualizado usando o box-plot.

QQ Plot dos resíduos estudentizados → os resíduos não tem distribuição aproximadamente normal.

c)

Modelo 2: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = -2{,}662 + 0{,}112 X_i ; \quad i = 1, 2, ..., 11$

R^2=98,9% e R^2ajust=98,8%

Intercepto: p-valor <0,001 < 0,05

Nº de tels: p-valor <0,001 < 0,05

Análise gráfica dos resíduos estudentizados e da distância de Cook → a observação 13 não é outlier, e sim uma observação altamente influente (D_{13}=7,903>1) segundo a distância de Cook.

Ainda há desvios da distribuição dos resíduos, relativamente à distribuição normal. A hipótese de homocedasticidade ainda não foi atendida.

d)

Modelo 3: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = -3{,}477 + 0{,}127 X_i ; \quad i = 1, 2, ..., 10$

R^2=94,8% e R^2ajust=94,1%

Intercepto: p-valor<0,001 < 0,05

Nº de tels: p-valor<0,001 < 0,05

As hipóteses de homocedasticidade e normalidade estão aparentemente satisfeitas. Além disso, a hipótese de linearidade está convenientemente atendida (*ver gráfico de dispersão entre X e Y*).

5ª Questão:

a) Todos os três gráficos revelam a existência de uma relação negativa (ou decrescente) entre: X_1 e Y (moderada), X_2 e Y (fraca) e X_3 e Y (moderada).

b)

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} = 332{,}111 - 1{,}546 X_{i1} - 1{,}425 X_{i2} - 2{,}237 X_{i3} ; \forall i = 1, 2, ..., 19$

Todos os parâmetros são significativos (p-valor < 0,001 < 0,05).

R^2ajust=94,6%

Não existe problema de multicolinearidade (VIF ≤10).

c) Os resíduos se encontram aleatoriamente distribuídos. Obs 12 é um *outlier*.

d) Valor crítico: 2p/n = 8/19 = 0,421 → Obs 1 é uma observação de alta alavancagem (alto potencial para influenciar o ajuste do modelo), e sua medida de alavancagem (h_{11}=0,430) está demasiadamente afastada das medidas das demais observações.

e) Todos os valores das distância de Cook, são menores do que 1 (D<1), mas as observações 2, 12, 14 e 18 apresentam valores de D bem superiores aos das demais observações, o que também pode ser comprovado por meio de um box-plot. Note que a obs 1, segundo a distância de Cook, não influencia substancialmente o ajuste do modelo.

f) Das quatro observações influentes identificadas na letra (e), apenas a obs 12 é considerada *outlier*.

g) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} = 334{,}253 - 1{,}571 X_{i1} - 1{,}419 X_{i2} - 1{,}929 X_{i3} ; \forall i = 1, 2, ..., 15$

Todos os parâmetros continuaram sendo significativos (p-valor < 0,001 < 0,05), indicando o mesmo sentido das relações. Apenas a estimativa de β_3 teve uma alteração mais expressiva.

R^2ajust=98,4% (melhora na qualidade do ajuste)

6ª Questão:

a)

Modelo 1: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 36{,}125 - 0{,}112 X_i ; \quad i = 1, 2, ..., 100$

R^2=28,8% e R^2ajust=28,1%

Intercepto: p-valor<0,001 < 0,05

% domicílios inadequados: p-valor<0,001 < 0,05

As hipóteses de homocedasticidade, independência e normalidade dos erros foram satisfeitas (*análise gráfica*).

b) Outliers: obs. 55, 61, 80, 78 e 97

Observações mais influentes: obs 90 (D_{90}=0,084) e 97 (D_{97}=0,072)

c)

Modelo 2: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 36{,}134 - 0{,}113 X_i$; $i = 1, 2, ..., 98$

R^2=30,6% e R^2ajust=29,8%

Intercepto: p-valor<0,001 < 0,05

% domicílios inadequados: p-valor<0,001 < 0,05

Não se verificou diferenças substancias nas estimativas dos parâmetros do modelo; e houve apenas uma ligeira melhora na qualidade do ajuste com a exclusão das duas observações mais influentes (note que a magnitude de D para todas as observações são baixas, embora essas duas observações tenham maiores valores de D em relação às demais). As hipóteses básicas do modelo continuam sendo satisfeitas (*análise gráfica*).

d) *Por conta do aluno!!!*

7ª Questão:

a) *Por conta do aluno !!!*

b) O gráfico parece sugerir que existe heterocedasticidade proporcional.

c) $\hat{Y}_i = 51{,}078 + 8{,}023 X_i$; $i = 1, 2, ..., 30$

Ambos os parâmetros são significativos (p-valor < 0,001 < 0,05).

R^2=95,4%

Teste de White: $w_{obs} = 7{,}08 > \chi^2_{0{,}05;2} = 5{,}991 \rightarrow$ existe heterocedasticidade.

<u>Regressão auxiliar:</u>

$e_i^2 = \hat{\gamma}_0 + \hat{\gamma}_1 X_{i1} + \hat{\gamma}_2 X_{i2}^2 = 29{,}221 - 3{,}936 X_{i1} + 0{,}544 X_{i2}^2$; $\forall\, i = 1, 2, ..., 30$

R^2=23,6%

d) $\hat{Y}_i^* = 7{,}951 + 51{,}927 X_i^*$; $i = 1, 2, ..., 30$ R^2=96,8%

No modelo transformado ambos os parâmetros são significativos (p-valor<0,001 < 0,05).

$$\frac{\hat{Y}_i}{X_i} = 7{,}951 + 51{,}927 \cdot \frac{1}{X_i}$$

$$\hat{Y}_i = 7{,}951 X_i + 51{,}927$$

$$\hat{Y}_i = 51{,}927 + 7{,}951 X_i \text{ ; } i = 1, 2, ..., 30$$