Hipótese RLS.3 (variação Amostral na variável explicativa)

Os resultados amostrais na x, a saber $\{x_i, i = 1, ..., n\}$ não são todos de mesmo valor.

Hipótese RLS.4 (Média Condicional Zero)

O erro u tem zero como valor esperado, quaisquer que sejam os valores das variáveis. Em outras palavras,

$$E(u|x) = 0.$$

Hipótese RLS.5 (Homoscedasticidade)

O erro u tem a mesma variância quaisquer que sejam os valores das variáveis explicativas. Em outras palavras:

$$Var(u|x) = \sigma^2.$$

PROBLEMAS

2.1 No modelo de regressão linear simples $y = \beta_0 + \beta_1 x + u$, suponha que $E(u) \neq 0$. Fazendo $\alpha_0 = E(u)$, mostre que o modelo pode sempre ser reescrito com a mesma inclinação, mas com um novo intercepto e erro, em que o novo erro tem um valor esperado zero.

2.2 A tabela seguinte contém as variáveis *supGPA* (nota média em curso superior nos Estados Unidos) e *ACT* (nota do teste de avaliação de conhecimentos para ingresso em curso superior nos Estados Unidos) com as notas hipotéticas de oito estudantes de curso superior. O *GPA* está baseada em uma escala de quatro pontos e foi arredondada para um dígito após o ponto decimal. A nota *ACT* baseia-se em uma escala de 36 pontos e foi arredondada para um número inteiro.

Estudante	*supGPA*	*ACT*
1	2,8	21
2	3,4	24
3	3,0	26
4	3,5	27
5	3,6	29
6	3,0	25
7	2,7	25
8	3,7	30

(i) Estime a relação entre *GPA* e *ACT* usando MQO; isto é, obtenha as estimativas de intercepto e de inclinação da equação

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 act.$$

Comente a direção da relação. O intercepto tem uma interpretação útil aqui? Explique. Qual deveria ser o valor previsto do *GPA* se a nota *ACT* aumentasse em cinco pontos?

(ii) Calcule os valores estimados e os resíduos de cada observação e verifique que a soma dos resíduos é (aproximadamente) zero.

(iii) Qual é o valor previsto do *GPA* quando $ACT = 20$?

(iv) Quanto da variação do *GPA* dos 8 estudantes é explicada pelo *ACT*? Explique.

2.3 Seja *filhos* o número de filhos de uma mulher, e *educ* os anos de educação da mulher. Um modelo simples que relaciona a fertilidade a anos de educação é

$$filhos = \beta_0 + \beta_1 educ + u,$$

em que u é um erro não observável.

(i) Que tipos de fatores estão contidos em u? É provável que eles estejam correlacionados com o nível de educação?

(ii) Uma análise de regressão simples mostrará o efeito *ceteris paribus* da educação sobre a fertilidade? Explique.

2.4 Suponha que você está interessado em estimar o efeito das horas gastas com um curso de preparação para o vestibular (*horas*) no total das notas do vestibular (*SAT*). A população são todos os pré-universitários graduados no ensino médio em determinado ano.

(i) Suponha que lhe tenha sido dado uma subvenção para executar um experimento controlado. Explique como você estruturaria o experimento de forma a estimar o efeito causal de *horas* no *SAT*.

(ii) Considere o caso mais realístico em que os alunos decidem quanto tempo gastarão com um curso de preparação, e você só pode fazer amostragens aleatórias de *SAT* e *horas* da população. Escreva o modelo populacional da seguinte forma

$$SAT = \beta_0 + \beta_1 horas = u$$

em que, como sempre, num modelo com um intercepto, podemos assumir $E(u) = 0$. Liste pelo menos dois fatores contidos na u. Existe a probabilidade de eles terem correlação negativa ou positiva com *horas*?

(iii) Na equação da parte (ii), qual deveria ser o sinal da β_1 se o curso de preparação for eficaz?

(iv) Na equação da parte (ii), qual é a interpretação da β_0?

2.5 Considere a função de poupança

$$poup = \beta_0 + \beta_1 rend + u,\ u = \sqrt{rend} \cdot e,$$

em que e é uma variável aleatória com $E(e) = 0$ e $Var(e) = \sigma_e^2$. Considere e independente de *rend*.

(i) Mostre que $E(u|rend) = 0$, de modo que a hipótese de média condicional zero (Hipótese RLS.4) é satisfeita. [*Sugestão*: se e é independente de *rend*, então $E(e|rend) = E(e)$.]

(ii) Mostre que $Var(u|rend) = \sigma_e^2 rend$, de modo que a hipótese de homoscedasticidade RLS.5 é violada. Em particular, a variância de *poup* aumenta com *rend*. [*Sugestão*: $Var(e/rend) = Var(e)$, se e e *rend* são independentes.]

(iii) Faça uma discussão que sustente a hipótese de que a variância da poupança aumenta com a renda da família.

2.6 Que $\hat{\beta}_0$ e $\hat{\beta}_1$ sejam os estimadores MQO do intercepto e inclinação, respectivamente, e que $\bar{u}$ seja a média amostral dos erros (não os resíduos!).

(i) Demonstre que $\hat{\beta}_1$ pode ser escrita como $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n w_i u_i$ em que $w_i = d_i/SQT_x$ e $d_i = x_i - \bar{x}$.

(ii) Use a parte (i), juntamente com $\sum_{i=1}^n w_i = 0$, para demonstrar que $\hat{\beta}_1$ e $\bar{u}$ são não correlacionadas. [*Sugestão*: Você está sendo solicitado a demonstrar que $E[\beta_1 - \beta_1) \cdot \bar{u}] = 0$.].

(iii) Demonstre que $\hat{\beta}_0$ pode ser escrita da seguinte forma $\hat{\beta}_0 = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}$.

(iv) Use as partes (ii) e (iii) para provar que $Var(\hat{\beta}_0) = \sigma^2/n + \sigma^2(\bar{x})^2/SQT_x$.

(v) Faça os cálculos para simplificar a expressão na parte (iv) para a equação (2.58). [*Sugestão*: $SQT_x/n = n^{-1}\sum_{i=1}^n x_i^2 - (\bar{x})^2$.]

2.7 Usando dados de casas vendidas em 1988 em Andover, Massachusetts [Kiel e McClain (1995)], a equação seguinte relaciona os preços das casas (*preço*) à distância de um incinerador de lixo recentemente construído (*dist*):

$$\widehat{\log(preço)} = 9{,}40 + 0{,}312 \log(dist)$$

$$n = 135,\ R^2 = 0{,}162.$$

(i) Interprete o coeficiente de $\log(dist)$. O sinal dessa estimativa é o que você esperava?

(ii) Você considera que a regressão simples oferece um estimador não viesado da elasticidade *ceteris paribus* de *preço* em relação a *dist*? (Pense sobre a decisão da cidade em localizar o incinerador.)

(iii) Quais outros fatores relativos a casas afetam seu preço? Eles poderiam estar correlacionados com a distância do incinerador?

2.8 (i) Sejam $\hat{\beta}_0$ e $\hat{\beta}_1$ o intercepto e a inclinação da regressão de y_i sobre x_i, usando n observações. Sejam c_1 e c_2 constantes, com $c_2 \neq 0$. Sejam $\tilde{\beta}_0$ e $\tilde{\beta}_1$ o intercepto e a inclinação da regressão de $c_1 y_i$ sobre $c_2 x_i$. Mostre que $\tilde{\beta}_1 = (c_1/c_2)\hat{\beta}_0$ e $\tilde{\beta}_0 = c_1 \hat{\beta}_0$, verificando as observações sobre as unidades de medida da Seção 2.4. [*Sugestão*: para obter $\tilde{\beta}_1$, insira as transformações de x e y em (2.19). Em seguida, use (2.17) para $\tilde{\beta}_0$, estando seguro de usar as transformações de x e y e a inclinação correta.]

(ii) Agora, sejam $\tilde{\beta}_0$ e $\tilde{\beta}_1$ os parâmetros estimados da regressão de $(c_1 + y_i)$ sobre $(c_2 + x_i)$ (sem qualquer restrição sobre c_1 ou c_2). Mostre que $\tilde{\beta}_1 = \hat{\beta}_1$ e $\tilde{\beta}_0 = \hat{\beta}_0 + c_1 - c_2 \hat{\beta}_1$.

(iii) Agora, sejam $\hat{\beta}_0$ e $\hat{\beta}_1$ as estimativas de MQO da regressão $\log(y_i)$ sobre x_i, para a qual devemos assumir $y_i > 0$ para todo i. Para $c_1 > 0$, sejam $\tilde{\beta}_0$ e $\tilde{\beta}_1$ o intercepto e a inclinação da regressão de $\log(c_1 y_i)$ sobre x_i. Mostre que $\tilde{\beta}_1 = \hat{\beta}_1$ e $\tilde{\beta}_0 = \log(c_1) + \hat{\beta}_0$.

(iv) Agora, assumindo que $x_i > 0$ para todo i, sejam $\tilde{\beta}_0$ e $\tilde{\beta}_1$ o intercepto e a inclinação da regressão de y_i sobre $\log(c_2 x_i)$. Como $\tilde{\beta}_0$ e $\tilde{\beta}_1$ comparam-se com o intercepto e a inclinação da regressão de y_i sobre $\log(x_i)$?

2.9 Na função de consumo linear

$$\widehat{cons} = \hat{\beta}_0 + \hat{\beta}_1 rend,$$

a *propensão marginal a consumir* PMgC (estimada) é simplesmente a inclinação $\hat{\beta}_1$, ao passo que a *propensão média a consumir* PmeC é $\widehat{cons}/rend = \hat{\beta}_0/rend + \hat{\beta}_1$. Usando as observações de renda e consumo anuais de 100 famílias (ambas medidas em dólares), obteve-se a seguinte equação:

$$\widehat{cons} = -124{,}84 + 0{,}853\ rend$$
$$n = 100,\ R^2 = 0{,}692.$$

(i) Interprete o intercepto dessa equação e comente seu sinal e magnitude.
(ii) Qual é o consumo previsto quando a renda familiar é US$ 30.000?
(iii) Com *rend* sobre o eixo de x, desenhe um gráfico da PMgC e da PmeC estimadas.

2.10 Considere o modelo de regressão simples padrão $y = \beta_0 + \beta_1 x + u$, sob as Hipóteses RLS.1 a RLS.4. Os estimadores usuais $\hat{\beta}_0$ e $\hat{\beta}_1$ são não viesados para seus respectivos parâmetros populacionais. Seja $\tilde{\beta}_1$ o estimador de β_1 obtido ao assumir que o intercepto é zero (veja a Seção 2.6).

(i) Encontre $E(\tilde{\beta}_1)$ em termos de x_i, β_0 e β_1. Verifique que $\tilde{\beta}_1$ é não viesado para β_1 quando o intercepto populacional (β_0) é zero. Há outros casos em que $\tilde{\beta}_1$ é não viesado?
(ii) Encontre a variância de $\tilde{\beta}_1$. [*Sugestão*: a variância não depende de β_0.]
(iii) Mostre que $\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1)$. [*Sugestão*: para qualquer amostra de dados, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i}^{n} (x_i - \bar{x})^2$, com a desigualdade estrita preponderando, a não ser que $\bar{x} = 0$.]
(iv) Comente a relação entre viés e variância, ao escolher entre $\hat{\beta}_1$ e $\tilde{\beta}_1$.

2.11 Os dados do arquivo BWGHT.RAW contém dados de nascimentos por mulheres nos Estados Unidos. As duas variáveis de interesse são: a variável dependente, peso dos recém-nascidos em onças* (*pesonas*), e a variável explicativa, número médio de cigarros que a mãe fumou por dia durante a gravidez (*cigs*). A seguinte regressão simples foi estimada usando dados de $n = 1.388$ nascimentos:

$$\widehat{pesonas} = 119{,}77 - 0{,}514\, cigs$$

(i) Qual é o peso de nascimento previsto quando $cigs = 0$? E quando $cigs = 20$ (um maço por dia)? Comente a diferença.
(ii) O modelo de regressão simples necessariamente captura uma relação causal entre o peso de nascimento da criança e os hábitos de fumar da mãe? Explique.
(iii) Para prever um peso de nascimento de 125 onças, qual deveria ser a magnitude de *cigs*? Comente.
(iv) Qual a fração de mulheres na amostra que não fumaram enquanto estiveram grávidas? Isso ajuda a reconciliar sua conclusão da parte (iii)?

APÊNDICE 2A

Minimizando a soma dos quadrados dos resíduos

Mostramos aqui que as estimativas de MQO $\hat{\beta}_0$ e $\hat{\beta}_1$ minimizam a soma dos quadrados dos resíduos, como afirmado na Seção 2.2. Formalmente, o problema é caracterizar as soluções $\hat{\beta}_0$ e $\hat{\beta}_1$ para o problema de minimização.

$$\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

em que b_0 e b_1 são argumentos *dummies* para o problema de otimização; para clareza, chame essa função $Q(b_0, b_1)$. De um resultado fundamental do cálculo multivariado (veja Apêndice A, disponível no site de Cengage), uma condição necessária para $\hat{\beta}_0$ e $\hat{\beta}_1$ resolverem o problema de minimização é que as derivadas parciais de $Q(b_0, b_1)$ em relação a b_0 e b_1 devem ser zero quando avaliadas com $\hat{\beta}_0, \hat{\beta}_1$: $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_0 = 0$ e $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_1 = 0$. Usando a regra da cadeia do cálculo, essas duas equações tornam-se

* 1 onça = 31,10 g (N.T.).

Hipótese RLM.3 (Colinearidade Imperfeita)

Na amostra (e, portanto, na população), nenhuma das variáveis independentes é constante e não existem relacionamentos *lineares exatos* entre as variáveis independentes.

Hipótese RLM.4 (Média Condicional Zero)

O erro u tem zero, como valor esperado, dados quaisquer valores das variáveis independentes. Em outras palavras:

$$\mathrm{E}(u|x_1, x_2, ..., x_k) = 0.$$

Hipótese RLM.5 (Homoscedasticidade)

O erro u tem a mesma variância, dados quaisquer valores das variáveis explicativas. Em outras palavras:

$$\mathrm{Var}(u|x_1, ..., x_k) = \sigma^2.$$

PROBLEMAS

3.1 Os dados do arquivo WAGE2.RAW, sobre homens que trabalham, foram utilizados para estimar a seguinte equação:

$$\widehat{educ} = 10{,}36 - 0{,}094\ irms + 0{,}131\ educm + 0{,}210\ educp$$
$$n = 722,\ R^2 = 0{,}214,$$

em que *educ* é anos de escolaridade formal, *irms* é o número de irmãos, *educm* é anos de escolaridade formal da mãe, e *educp* é anos de escolaridade formal do pai.

(i) *irms* tem o efeito esperado? Explique. Mantendo *educm* e *educp* fixos, em quanto deveria *irms* aumentar para reduzir os anos previstos da educação formal em um ano? (Uma resposta incompleta é aceitável aqui.)
(ii) Discuta a interpretação do coeficiente de *educm*.
(iii) Suponha que o Homem A não tenha irmãos, e sua mãe e seu pai tenham, cada um, 12 anos de educação formal. Suponha também que o Homem B não tenha irmãos, e sua mãe e seu pai tenham, cada um, 16 anos de educação formal. Qual é a diferença prevista em anos de educação formal entre B e A?

3.2 Usando os dados do arquivo GPA2.RAW sobre 4.137 estudantes de curso superior nos Estados Unidos, estimou-se a seguinte equação por MQO:

$$\widehat{supGPA} = 1{,}392 - 0{,}0135\ emperc + 0{,}00148\ SAT$$
$$n = 4.137,\ R^2 = 0{,}273,$$

em que *supGPA* é mensurada em uma escala de quatro pontos, *emperc* é o percentual da turma de formados do ensino médio (definido de modo que, por exemplo, *emperc* = 5 significa os 5% *melho-*

res da sala), e *SAT* é uma nota média ponderada de matemática e habilidade verbal do estudante para ingresso em curso superior.

(i) Por que faz sentido que o coeficiente de *emperc* seja negativo?
(ii) Qual é o valor previsto de *supGPA* quando *emperc* = 20 e *SAT* = 1.050?
(iii) Suponha que dois alunos do ensino médio, A e B, estejam no mesmo percentual no ensino médio, mas a nota *SAT* do Estudante A foi 140 pontos maior (cerca de um desvio-padrão na amostra). Qual é a diferença prevista em *supGPA* para esses dois estudantes? A diferença é grande?
(iv) Mantendo *emperc* fixo, que diferença na nota *SAT* levaria a uma diferença prevista de *supGPA* de 0,50? Comente sua resposta.

3.3 O salário inicial (mediano) para recém-formados em direito é determinado pela equação

$$\log(salárioim) = \beta_0 + \beta_1 lSAT + \beta_2 supGPA + \beta_3 \log(volbib) + \beta_4 \log(custo) + \beta_5 rank + u,$$

em que *lSAT* é a nota mediana do *lSAT* (nota de ingresso no curso de direito) dos recém-formados, *supGPA* é a nota mediana dos recém-formados nas disciplinas do curso de direito, *volbib* é o número de volumes da biblioteca da escola de direito, *custo* é o custo anual da escola de direito e *rank* é a classificação da escola de direito (com *rank* = 1 sendo o melhor posto de classificação).

(i) Explique a razão de esperarmos $\beta_5 \leq 0$.
(ii) Quais são os sinais que você espera para os outros parâmetros de inclinação? Justifique sua resposta.
(iii) Utilizando os dados do arquivo LAWSCH85.RAW, a equação estimada é

$$\widehat{\log(salário)} = 8,34 + 0,0047\ lSAT + 0,248\ supGPA + 0,095 \log(volbib) + 0,038 \log(custo) - 0,0033\ rank$$
$$n = 136,\ R^2 = 0,842.$$

Qual é a diferença *ceteris paribus* prevista no salário para as escolas com um *supGPA* mediano diferente em um ponto? (Descreva sua resposta em percentual.)

(iv) Interprete o coeficiente da variável log(*volbib*).
(v) Você diria que é melhor frequentar uma escola de direito que tem uma classificação melhor? Qual é a diferença no salário inicial esperado para uma escola que tem uma classificação igual a 20?

3.4 O modelo seguinte é uma versão simplificada do modelo de regressão múltipla usado por Biddle e Hamermesh (1990) para estudar a escolha entre o tempo gasto dormindo e trabalhando e para observar outros fatores que afetam o sono:

$$dormir = \beta_0 + \beta_1 trabtot + \beta_2 educ + \beta_3 idade + u,$$

em que *dormir* e *trabtot* (trabalho total) são mensurados em minutos por semana e *educ* e *idade* são mensurados em anos. (Veja também a seção Exercícios em Computador 2.3, no site da Cengage.)
(i) Se os adultos escolhem entre dormir e trabalhar, qual é o sinal de β_1?
(ii) Que sinais você espera que β_2 e β_3 terão?
(iii) Usando os dados do arquivo SLEEP75.RAW, a equação estimada é

$$\widehat{dormir} = 3.638,25 - 0,148\ trabtot - 11,13\ educ + 2,20\ idade$$
$$n = 706,\ R^2 = 0,113.$$

Se alguém trabalha cinco horas a mais por semana, qual é a queda, em minutos, no valor esperado de *dormir*? Esse valor representa uma escolha grande?

(iv) Discuta o sinal e a magnitude do coeficiente de *educ*.

(v) Você diria que *trabtot*, *educ* e *idade* explicam muito da variação de *dormir*? Quais outros fatores poderiam afetar o tempo gasto dormindo? É provável que eles sejam correlacionados com *trabtot*?

3.5 Considere o modelo de regressão múltipla contendo três variáveis independentes, sob as Hipóteses RLM.1 a RLM.4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

Você está interessado em estimar a soma dos parâmetros de x_1 e x_2; chame-a de $\theta_1 = \beta_1 + \beta_2$.
(i) Mostre que $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ é um estimador não viesado de θ_1.
(ii) Encontre $\text{Var}(\hat{\theta}_1)$ em termos de $\text{Var}(\hat{\beta}_1)$, $\text{Var}(\hat{\beta}_2)$, e $\text{Corr}(\hat{\beta}_1, \hat{\beta}_2)$.

3.6 Em um estudo que relaciona a nota média em curso superior (*supGPA*) ao tempo gasto em várias atividades, você distribui uma pesquisa para vários estudantes. Os estudantes devem responder quantas horas eles despendem, em cada semana, em quatro atividades: estudo, sono, trabalho e lazer. Toda atividade é colocada em uma das quatro categorias, de modo que, para cada estudante, a soma das horas nas quatro atividades deve ser igual a 168.
(i) No modelo

$$supGPA = \beta_0 + \beta_1 estudar + \beta_2 dormir + \beta_3 trabalhar + \beta_4 lazer + u,$$

faz sentido manter *dormir*, *trabalhar* e *lazer* fixos, enquanto *estudar* varia?

(ii) Explique a razão de esse modelo violar a Hipótese RLM.3.
(iii) Como você poderia reformular o modelo, de modo que seus parâmetros tivessem uma interpretação útil e ele satisfizesse a Hipótese RLM.4?

3.7 Suponha que a produtividade média do trabalhador da indústria (*prodmed*) dependa de dois fatores — horas médias de treinamento do trabalhador (*treinmed*) e aptidão média do trabalhador (*aptidmed*):

$$prodmed = \beta_0 + \beta_1 treinmed + \beta_2 aptidmed + u.$$

Suponha que essa equação satisfaça as hipóteses de Gauss-Markov. Se um subsídio foi dado às empresas cujos trabalhadores têm uma aptidão menor do que a média, de modo que *treinmed* e *aptidmed* sejam negativamente correlacionados, qual é o provável viés em $\tilde{\beta}_1$ obtido da regressão simples de *prodmed* sobre *treinmed*?

3.8 Quais dos seguintes itens podem fazer com que os estimadores de MQO sejam viesados?
(i) Heteroscedasticidade.
(ii) Omitir uma variável importante.
(iii) Um coeficiente de correlação amostral de 0,95 entre duas variáveis independentes incluídas no modelo.

3.9 Suponha que você tenha interesse em estimar o relacionamento *ceteris paribus* entre y e x_1. Para esse propósito você pode coligir dados em duas variáveis de controle, x_2 e x_3. (Para melhor clareza, você pode entender y como uma nota do exame final, x_1 como frequência às aulas, x_2 como a nota de média graduação até o semestre anterior, e x_3 como uma nota de teste de aptidão acadêmica ou de teste de avaliação.) Seja $\tilde{\beta}_1$ a estimativa da regressão simples de y sobre x_1 e seja $\hat{\beta}_1$ a estimativa de regressão múltipla de y sobre x_1, x_2, x_3.

(i) Se x_1 for altamente correlacionada com x_2 e x_3 na amostra e x_2 e x_3 tiverem grandes efeitos parciais na y, você antecipa que $\tilde{\beta}_1$ e $\hat{\beta}_1$ sejam semelhantes ou muito diferentes? Explique.

(ii) Se x_1 for quase não correlacionada com x_2 e x_3, mas x_2 e x_3 forem altamente correlacionadas, as $\tilde{\beta}_1$ e $\hat{\beta}_1$ tenderão a ser semelhantes ou muito diferentes? Explique.

(iii) Se x_1 for altamente correlacionada com x_2 e x_3 e x_2 e x_3 tiverem pequenos efeitos parciais em y, você anteciparia que $ep(\tilde{\beta}_1)$ ou $ep(\hat{\beta}_1)$ será menor? Explique.

(iv) Se x_1 for quase não correlacionada com x_2 e x_3, x_2 e x_3 tiver grandes efeitos parciais em y, e x_2 e x_3 forem altamente correlacionadas, você anteciparia que $ep(\tilde{\beta}_1)$ ou $ep(\hat{\beta}_1)$será menor? Explique.

3.10 Suponha que o modelo populacional que determina y seja:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

e esse modelo satisfaz as hipóteses de Gauss-Markov. Entretanto, estimamos o modelo que omite x_3. Sejam, $\tilde{\beta}_0$, $\tilde{\beta}_1$ e $\tilde{\beta}_2$ os estimadores de MQO da regressão de y sobre x_1 e x_2. Mostre que o valor esperado de β_1 (dados os valores das variáveis independentes da amostra) é

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2},$$

em que os $\hat{r}_{i1}$ são os resíduos de MQO da regressão de x_1 sobre x_2. [*Sugestão*: a fórmula de $\tilde{\beta}_1$ é proveniente da equação (3.22). Coloque $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$ nessa equação. Após alguma álgebra, aplique o operador expectativa, tratando x_{i3} e $\hat{r}_{i1}$ como não aleatórios.]

3.11 A equação seguinte descreve o preço mediano das residências de uma comunidade em termos da quantidade de poluição (*oxn*, de óxido nitroso) e do número médio de cômodos nas residências da comunidade (*comods*):

$$\log(preço) = \beta_0 + \beta_1 \log(oxn) + \beta_2 comods + u.$$

(i) Quais são os prováveis sinais de β_1 e β_2? Qual é a interpretação de β_1? Explique.

(ii) Por que *oxn* [ou, mais precisamente, log(*oxn*)] e *comods* deveriam ser negativamente correlacionados? Se esse é o caso, a regressão simples de log(*preço*) sobre log(*oxn*) produz um estimador viesado para cima ou para baixo de β_1?

(iii) Utilizando os dados do arquivo HPRICE2.RAW foram estimadas as seguintes equações:

$$\widehat{\log(preço)} = 11{,}71 - 1{,}043 \log(oxn),\ n = 506,\ R^2 = 0{,}264.$$

$$\widehat{\log(preço)} = 9{,}23 - 0{,}718 \log(oxn) + 0{,}306\ comods,\ n = 506,\ R^2 = 0{,}514.$$

A relação entre as estimativas da elasticidade do *preço* das regressões simples e múltipla é a que você previu, tomando como base sua resposta na parte (ii)? Pode-se dizer que $-0{,}718$ está claramente mais próximo da elasticidade verdadeira que $-1{,}043$?

3.12 Leia os itens abaixo e faça o que se pede.

(i) Considere o modelo de regressão simples $y = \beta_0 + \beta_1 x + u$, sob as primeiras quatro hipóteses de Gauss-Markov. Para alguma função $g(x)$, por exemplo, $g(x) = x^2$ ou $g(x) = \log(1 + x^2)$, defina $z_i = g(x_i)$. Defina um estimador de inclinação como

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n} (z_i - \bar{z}) y_i\right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z}) x_i\right).$$

Mostre que $\tilde{\beta}_1$ é linear e não viesado. Lembre-se: como $\mathrm{E}(u|x) = 0$, você pode tratar tanto x_i como z_i como não aleatórios em sua derivação.

(ii) Acrescente a hipótese de homoscedasticidade, RLM.5. Mostre que

$$Var(\tilde{\beta}_1) = \sigma^2 \left(\sum_{i=1}^{n} (z_i - \bar{z})^2\right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z}) x_i\right)^2.$$

(iii) Mostre diretamente que, sob as hipóteses de Gauss-Markov, $\mathrm{Var}(\hat{\beta}_1) \leq \mathrm{Var}(\tilde{\beta}_1)$, em que $\hat{\beta}_1$ é o estimador de MQO. [*Sugestão*: a desigualdade de Cauchy-Schwartz do Apêndice B (disponível no site da Cangage) implica que

$$\left(n^{-1} \sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})\right)^2 \leq \left(n^{-1} \sum_{i=1}^{n} (z_i - \bar{z})^2\right)\left(n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2\right);$$

Observe que podemos retirar da covariância amostral.]

3.13 A seguinte equação representa os efeitos das receitas totais de impostos sobre o crescimento subsequente do emprego para a população de municípios dos Estados Unidos:

$$cresc = \beta_0 + \beta_1 parc_{\mathrm{p}} + \beta_2 parc_{\mathrm{r}} + \beta_3 parc_{\mathrm{v}} + outros\ fatores,$$

em que *cresc* é a variação percentual do emprego de 1980 a 1990, enquanto o total das receitas de impostos tem a seguinte distribuição: $parc_{\mathrm{p}}$ é a parcela dos impostos sobre a propriedade, $parc_{\mathrm{r}}$ é a parcela das receitas de impostos sobre a renda, e $parc_{\mathrm{v}}$ é a parcela das receitas de impostos sobre as vendas. Todas essas variáveis estão mensuradas em 1980. A parcela omitida, $parc_{\mathrm{t}}$, inclui taxas e impostos variados. Por definição, as quatro parcelas somam um. Outros fatores incluiriam despesas com educação, infraestrutura, e assim por diante (todos mensurados em 1980).

(i) Por que devemos omitir uma das variáveis de parcela de impostos da equação?

(ii) Dê uma interpretação cuidadosa de β_1.