Problemas

1. Utilizando os dados contidos no arquivo SLEEP75 (veja também o Problema 3.3), obtemos a equação estimada

$$\widehat{sleep} = 3.840{,}83 - 0{,}163\ totwrk - 11{,}71\ educ - 8{,}70\ age$$
$$(235{,}11)\ (0{,}018)\qquad (5{,}86)\qquad (11{,}21)$$
$$+ 0{,}128\ age^2 + 87{,}75\ male$$
$$(0{,}134)\qquad (34{,}33)$$
$$n = 706,\ R^2 = 0{,}123,\ \bar{R}^2 = 0{,}117.$$

A variável *sleep* é o total de minutos gastos por semana dormindo durante a noite, *totwrk* é o total de minutos semanais gastos trabalhando, *educ* e *age* são medidas em anos e *male* é uma variável *dummy* de gênero.

(i) Supondo todos os outros fatores iguais, existe evidência de que os homens dormam mais que as mulheres? O quanto essa evidência é forte?

(ii) Existe uma relação de substituição estatisticamente significante entre trabalhar e dormir? Qual é a relação de substituição estimada?

(iii) Que outras regressões você precisa executar para testar a hipótese nula de que, mantendo fixos os outros fatores, a idade não tem efeito sobre dormir?

2. As seguintes equações foram estimadas utilizando os dados contidos no arquivo BWGHT:

$$\widehat{\log(bwght)} = 4{,}66 - 0{,}0044\ cigs + 0{,}0093\ \log(faminc) + 0{,}016\ parity$$
$$(0{,}22)\ (0{,}0009)\qquad (0{,}0059)\qquad (0{,}006)$$
$$+ 0{,}027\ male + 0{,}055\ white$$
$$(0{,}010)\qquad (0{,}013)$$
$$n = 1.388,\ R^2 = 0{,}0472$$

e

$$\widehat{\log(bwght)} = 4{,}65 - 0{,}0052\ cigs + 0{,}0110\ \log(faminc) + 0{,}017\ parity$$
$$(0{,}38)\ (0{,}0010)\qquad (0{,}0085)\qquad (0{,}006)$$
$$+ 0{,}034\ male + 0{,}045\ white - 0{,}0030\ motheduc + 0{,}0032\ fatheduc$$
$$(0{,}011)\qquad (0{,}015)\qquad (0{,}0030)\qquad (0{,}0026)$$
$$n = 1.191,\ R^2 = 0{,}0493.$$

As variáveis são definidas como no Exemplo 4.9, mas adicionamos uma variável *dummy* para o caso de a criança ser do sexo masculino e uma variável *dummy* que indica se a criança é classificada como branca.

(i) Na primeira equação, interprete o coeficiente da variável *cigs*. Particularmente, qual é o efeito no peso dos recém-nascidos se a mãe fumar dez ou mais cigarros por dia?

(ii) Quanto se espera que um recém-nascido branco pese a mais que uma criança não branca, mantendo fixos todos os outros fatores na primeira equação? A diferença é estatisticamente significante?

(iii) Comente sobre o efeito estimado e a significância estatística de *motheduc*.

(iv) Com a informação dada, por que você não terá condições de calcular a estatística *F* da significância conjunta de *motheduc* e *fatheduc*? O que você teria de fazer para calcular a estatística *F*?

3. Utilizando os dados contidos no arquivo GPA2, a seguinte equação foi estimada:

$$\widehat{sat} = 1.028{,}10 + 19{,}30\,hsize - 2{,}19\,hsize^2 - 45{,}09\,female$$
$$\quad (6{,}29) \quad (3{,}83) \quad (0{,}53) \quad (4{,}29)$$
$$- 169{,}81\,black + 62{,}31\,female{\cdot}black$$
$$\quad (12{,}71) \quad (18{,}15)$$
$$n = 4.137,\ R^2 = 0{,}0858.$$

A variável *sat* é a pontuação combinada de matemática e habilidade verbal do estudante para ingresso em curso superior (SAT); *hsize* é o tamanho da classe do aluno no ensino médio, em centenas; *female* é uma variável *dummy* para gênero; *black* é uma variável *dummy* da raça igual a um para negros, e zero, caso contrário.

(i) Existe forte evidência de que $hsize^2$ deva ser incluída no modelo? Desta equação, qual é o tamanho ótimo da classe no ensino médio?

(ii) Mantendo fixo *hsize*, qual é a diferença estimada na nota SAT entre mulheres não negras e homens não negros? O quanto é estatisticamente significante essa diferença estimada?

(iii) Qual é a diferença estimada na *sat* entre homens não negros e homens negros? Teste a hipótese nula de que não há diferença entre suas notas, contra a hipótese alternativa de que existe uma diferença.

(iv) Qual é a diferença estimada na nota *sat* entre mulheres negras e mulheres não negras? O que você necessitaria fazer para verificar se a diferença é estatisticamente significante?

4. Uma equação que explica os salários dos diretores-executivos é

$$\widehat{\log(salary)} = 4{,}59 + 0{,}257\log(sales) + 0{,}011\,roe + 0{,}158\,finance$$
$$\quad (0{,}30) \quad (0{,}032) \quad (0{,}004) \quad (0{,}089)$$
$$+ 0{,}181\,consprod - 0{,}283\,utility$$
$$\quad (0{,}085) \quad (0{,}099)$$
$$n = 209,\ R^2 = 0{,}357.$$

Os dados usados estão no arquivo CEOSAL1, no qual *finance*, *consprod* e *utility* são variáveis binárias que indicam as empresas financeiras, de produtos de consumo e de serviços públicos. O ramo de atividade omitido foi o de transportes.

(i) Calcule a diferença percentual aproximada no salário estimado entre os setores de serviços públicos e de transportes, mantendo fixos *sales* e *roe*. A diferença é estatisticamente significante ao nível de 1%?

(ii) Use a equação (7.10) para obter a diferença percentual exata no salário estimado entre os setores de serviços públicos e de transportes e compare-a com a resposta obtida no item (i).

(iii) Qual é a diferença percentual aproximada no salário estimado entre as atividades de produtos de consumo e financeiros? Escreva uma equação que possibilite verificar se a diferença é estatisticamente significante.

5. No Exemplo 7.2 defina *noPC* como uma variável *dummy* igual a um se o aluno não possuir um PC, e zero caso contrário.
   (i) Se *noPC* for usada no lugar de *PC* na equação (7.6), o que acontece com o intercepto na equação estimada? Qual será o coeficiente de *noPC*? (*Dica*: Escreva $PC = 1 - noPC$ e agregue isso na equação $\widehat{colGPA} = \hat{\beta}_0 + \hat{\delta}_0 PC + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT$.)
   (ii) O que acontecerá com o *R*-quadrado se *noPC* for usado em lugar de *PC*?
   (iii) As variáveis *PC* e *noPC* deveriam ser incluídas como variáveis independentes no modelo? Explique.
6. Para testar a eficiência de um programa de treinamento de pessoal sobre os subsequentes salários dos trabalhadores, especificamos o modelo

$$\log(wage) = \beta_0 + \beta_1 trein + \beta_2 educ + \beta_3 exper + u,$$

   em que *trein* é uma variável binária igual à unidade se um trabalhador participou do programa. Pense no termo de erro *u* como contendo a aptidão não observada do trabalhador. Se trabalhadores menos aptos tiverem maior oportunidade de ser selecionados para o programa, e você usar uma análise MQO, o que você pode dizer sobre o provável viés no estimador MQO de β_1? (*Dica*: Consulte o Capítulo 3.)
7. No exemplo da equação (7.29), suponha que definamos *outlf* como um se a mulher estiver fora da força de trabalho, e zero, caso contrário.
   (i) Se fizermos a regressão de *outlf* sobre todas as variáveis independentes na equação (7.29), o que acontecerá com as estimativas do intercepto e da inclinação? (Dica: $inlf = 1 - outlf$. Agregue essa expressão na equação populacional $inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \ldots$ e reorganize.)
   (ii) O que acontecerá com os erros padrão das estimativas do intercepto e da inclinação?
   (iii) O que acontecerá com o *R*-quadrado?
8. Suponha que você colete dados de uma pesquisa sobre salários, educação, experiência e gênero. Além disso, você solicita informações sobre o uso de maconha. A pergunta original é: "Em quantas ocasiões distintas, no mês passado, você fumou maconha?".
   (i) Escreva uma equação que permita a você estimar os efeitos do uso de maconha sobre os salários com todos os outros fatores controlados. Você deve ter condições de fazer declarações do tipo: "Estima-se que fumar maconha cinco vezes ou mais por mês altera os salários em *x*%".
   (ii) Escreva um modelo que permita verificar se o uso de drogas tem efeitos diferentes sobre os salários dos homens e das mulheres. Como você verificaria que não existem diferenças nos efeitos do uso de drogas nos homens e nas mulheres?
   (iii) Suponha que você considere ser melhor avaliar o uso de maconha colocando as pessoas em uma de quatro categorias: não usuário, usuário leve (um a cinco vezes por mês), usuário moderado (seis a dez vezes por mês), e usuário inveterado (mais de dez vezes por mês). Agora escreva um modelo que permita estimar os efeitos da maconha sobre os salários.
   (iv) Usando o modelo do item (iii), explique em detalhes como testar a hipótese nula de que o uso de maconha não tem efeito sobre o salário. Seja bastante específico e inclua uma relação cuidadosa de graus de liberdade.
   (v) Quais são alguns dos problemas potenciais de procurar inferência causal utilizando os dados da pesquisa que você coletou?

9. Que d seja uma variável *dummy* (binária) e que z seja uma variável quantitativa. Considere o modelo

$$y = \beta_0 + \delta_0 d + \beta_1 z + \delta_1 d \cdot z + u;$$

esta é uma versão geral de um modelo com uma interação entre uma variável *dummy* e uma quantitativa. [Um exemplo está na equação (7.17).]

(i) Como isto não alterará nada importante, defina o erro com valor zero, $u = 0$. Então, quando $d = 0$ podemos escrever o relacionamento entre y e z como a função $f_0(z) = \beta_0 + \beta_1 z$. Escreva a mesma relação quando $d = 1$, em que você deve usar $f_1(z)$ no lado esquerdo para denotar a função linear de z.

(ii) Considerando $\delta_1 \neq 0$ (o que significa que as duas não são paralelas), demonstre que o valor de z^* de tal forma que $f_0(z^*) = f_1(z^*)$ é $z^* = -\delta_0/\delta_1$. Este é o ponto no qual as duas linhas se cruzam [como na Figura 7.2(b)]. Demonstre que z^* será positivo se, e somente se, δ_0 e δ_1 tiverem sinais opostos.

(iii) Usando os dados contidos no arquivo TWOYEAR, a seguinte equação poderá ser estimada:

$$\widehat{\log(wage)} = \underset{}{2{,}289} - \underset{(0{,}011)}{0{,}357}\,female + \underset{(0{,}003)}{0{,}50}\,totcoll + \underset{(0{,}005)}{0{,}030}\,female \cdot totcoll$$
$$n = 6{,}763,\ R^2 = 0{,}202$$

em que todos os coeficientes e erros padrão foram arredondados com três casas decimais. Usando esta equação, encontre o valor de *totcoll* de tal forma que os valores previstos de log(*wage*) sejam os mesmos tanto para homens como para mulheres.

(iv) Com base na equação do item (iii), as mulheres podem, de forma realística, obter educação superior de modo que seus ganhos alcancem os dos homens? Explique.

10. Para uma criança i que mora em determinada região de ensino, defina $voucher_i$ como uma variável *dummy* igual a um se a criança foi selecionada para participar de um programa de bolsas de estudos em uma escola, e defina $score_i$ como a nota da criança em um exame padronizado subsequente. Suponha que a variável de participação, $voucher_i$, seja completamente aleatorizada para que ela seja independente tanto dos fatores observados quanto dos não observados que possam afetar a nota do teste de avaliação.

(i) Se você executar uma regressão simples de $score_i$ sobre $voucher_i$ usando uma amostra aleatória de tamanho n, o estimador de MQO produzirá um estimador não viesado do efeito do programa de bolsas de estudos?

(ii) Suponha que você possa coletar informações adicionais de perfis familiares tais como renda familiar, estrutura familiar (por exemplo, se a criança mora com os dois pais) e nível de escolaridade dos pais. Você precisará controlar esses fatores para obter um estimador não viesado dos efeitos do programa de bolsas de estudos? Explique.

(iii) Por que você precisará incluir as variáveis de perfis familiares na regressão? Existe uma situação em que você não incluiria as variáveis de perfis familiares?

11. As equações a seguir foram estimadas usando os dados do arquivo ECONMATH, com erros padrão registrados abaixo dos coeficientes. A nota média da classe, medida como porcentagem, é de cerca de 72,2; exatamente 50% dos estudantes são do sexo masculino; e a média de *colgpa* (nota média no início do semestre) é de cerca de 2,81.

$$\widehat{score} = 32{,}31 + 14{,}32\, colgpa$$
$$(2{,}00)\quad (0{,}70)$$
$$n = 856,\ R^2 = 0{,}329,\ \bar{R}^2 = 0{,}328$$

$$\widehat{score} = 29{,}66 + 3{,}83\, male + 14{,}57\, colgpa$$
$$(2{,}04)\ (0{,}74)\qquad (0{,}69)$$
$$n = 856,\ R^2 = 0{,}349,\ \bar{R}^2 = 0{,}348.$$

$$\widehat{score} = 30{,}36 + 2{,}47\, male + 14{,}33\, colgpa + 0{,}479\, male \cdot colgpa$$
$$(2{,}86)\ (3{,}96)\qquad (0{,}98)\qquad (1{,}383)$$
$$n = 856,\ R^2 = 0{,}349,\ \bar{R}^2 = 0{,}347.$$

$$\widehat{score} = 30{,}36 + 3{,}82\, male + 14{,}33\, colgpa + 0{,}479\, male \cdot (colgpa - 2{,}81)$$
$$(2{,}86)\ (0{,}74)\qquad (0{,}98)\qquad (1{,}383)$$
$$n = 856,\ R^2 = 0{,}349,\ \bar{R}^2 = 0{,}347.$$

(i) Interprete o coeficiente sobre *male* na segunda equação e construa um intervalo de confiança de 95% para β_{male}. Este intervalo de confiança exclui zero?

(ii) Na segunda equação, por que a estimativa de *male* é tão imprecisa? Devemos agora concluir que não existem diferenças de gênero em nota depois de controlar *colgpa*? [Dica: Você pode querer calcular a estatística F da hipótese nula de que não há diferença de gênero no modelo com a interação.]

(iii) Comparado com a terceira equação, como pode o coeficiente de *male* na última equação ser tão mais próximo daquele da segunda equação e tão precisamente estimado?

Exercícios em computador

C1 Use os dados do arquivo GPA1 neste exercício.

(i) Adicione as variáveis *mothcoll* e *fathcoll* à equação estimada em (7.6) e registre os resultados na forma usual. O que acontece com o efeito estimado da posse de computadores? A variável *PC* ainda é estatisticamente significativa?

(ii) Teste se há significância conjunta de *mothcoll* e *fathcoll* na equação do item (i) e certifique-se de registrar o *p*-valor.

(iii) Adicione *ACT* ao modelo do item (i) e decida se essa generalização é necessária.

C2 Use os dados do arquivo WAGE2 para este exercício.

(i) Estime o modelo

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

e registre os resultados na forma usual. Mantendo os outros fatores fixos, qual é a diferença aproximada no salário mensal entre negros (*black*) e não negros? Essa diferença é estatisticamente significativa?

(ii) Adicione as variáveis exper^2 e *tenure*^{2} à equação e mostre que eles são conjuntamente insignificantes mesmo no nível de 20%.

(iii) Amplie o modelo original para permitir que o retorno à educação dependa da raça e teste se o retorno à educação realmente depende da raça.

(iv) Novamente, comece com o modelo original, mas permita que os salários sejam diferentes entre quatro grupos de pessoas: casados e negros, casados e não negros, solteiros e negros e solteiros e não negros. Qual é o diferencial de salário estimado entre casados negros e casados não negros?

C3 Um modelo que permite que o salário de um jogador da liga principal de beisebol varie por posição é

$$\begin{aligned}\log(salary) = {} & \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr \\ & + \beta_5 rbisyr + \beta_6 runsyr + \beta_7 fldperc + \beta_8 allstar \\ & + \beta_9 frstbase + \beta_{10} scndbase + \beta_{11} thrdbase + \beta_{12} shrtstop \\ & + \beta_{13} catcher + u,\end{aligned}$$

onde os defensores (*outfield*) são o grupo-base.

(i) Declare a hipótese nula de que, controlados os outros fatores, receptores e defensores ganham, em média, a mesma quantia. Teste essa hipótese usando os dados do arquivo MLB1 e comente sobre a dimensão do diferencial estimado no salário.

(ii) Declare e teste a hipótese nula de que não existe diferença no salário médio entre as posições uma vez que outros fatores tenham sido controlados.

(iii) Os resultados dos itens (i) e (ii) são coerentes? Se não, explique o que está acontecendo.

C4 Use os dados do arquivo GPA2 para resolver este exercício.

(i) Considere a equação

$$\begin{aligned}colgpa = {} & \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat \\ & + \beta_5 female + \beta_6 athlete + u,\end{aligned}$$

onde *colgpa* é a nota média cumulativa na faculdade; *hsize* é o tamanho da classe de graduação no ensino médio, em centenas; *hsperc* é o percentual acadêmico na classe de graduação; sat é a pontuação SAT combinada; *female* é uma variável binária de gênero; e *athlete* é uma variável binária, que vale um para estudantes atletas. Quais são suas expectativas para os coeficientes dessa equação? Sobre quais deles você está inseguro?

(ii) Estime a equação do item (i) e registre os resultados na forma usual. Qual é o diferencial de *colgpa* estimado entre atletas e não atletas? Ele é estatisticamente significativo?

(iii) Retire *sat* do modelo e reestime a equação. Agora, qual é o efeito estimado de ser um atleta? Discuta por que a estimativa é diferente daquela obtida no item (ii).

(iv) No modelo do item (i), permita que o efeito de ser um atleta varie por gênero e teste a hipótese nula de que não há diferença *ceteris paribus* entre mulheres atletas e mulheres não atletas.

(iv) O efeito de *sat* sobre *colgpa* varia por gênero? Justifique sua resposta.

C5 No problema 2 do Capítulo 4, adicionamos o retorno sobre o capital das empresas, *ros*, a um modelo que explicava o salário dos CEOs; *ros* acabou sendo insignificante.

Agora, defina uma variável *dummy*, *rosneg*, que é igual a um se $ros < 0$ e igual a zero se $ros \geq 0$. Use o arquivo CEOSAL1 para estimar o modelo

$$\log(wage) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 rosneg + u$$

Discuta a interpretação e a significância estatística de $\hat{\beta}_3$.

C6 Use os dados do arquivo SLEEP75 para este exercício. A equação de interesse é

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u.$$

(i) Estime essa equação separadamente para homens e mulheres e registre os resultados na forma usual. Existem diferenças notáveis nas duas equações estimadas?

(ii) Calcule o teste de Chow para igualdade dos parâmetros na equação do sono para homens e mulheres. Use a forma do teste que adiciona *male* e os termos de interação *male·totwrk*, ..., *male·yngkid* e use o conjunto total de observações. Quais são os *gl* relevantes para o teste? Você deveria rejeitar a hipótese nula a um nível de 5%?

(iii) Agora permita um intercepto diferente para homens e mulheres e determine se os termos da interação envolvendo *male* são conjuntamente significativos.

(iv) Dados os resultados dos itens (ii) e (iii), qual seria seu modelo final?

C7 Use os dados do arquivo WAGE1 neste exercício.

(i) Use a equação (7.18) para estimar o diferencial de gênero quando *educ* =12,5. Compare isso com o diferencial estimado quando *educ* = 0.

(ii) Faça a regressão usada para obter a equação (7.19), mas com *female·*(*educ* - 12,5) substituindo *female·educ*. Como você interpreta o coeficiente sobre *female* agora?

(iii) O coeficiente sobre *female* do item (ii) é estatisticamente significativo? Compare isso com a equação (7.18) e comente.

C8 Use os dados do arquivo LOANAPP para este exercício. A variável binária a ser explicada é *approve*, que é igual a um se um empréstimo hipotecário para um indivíduo for aprovado. A principal variável explicativa é *white*, uma variável dummy igual a um se o solicitante for branco. Os outros solicitantes do conjunto de dados são negros e hispânicos.

Para testar se há discriminação no mercado de empréstimos hipotecários, um modelo de probabilidade linear pode ser usado:

$$approve = \beta_0 + \beta_1 white + outros\ fatores.$$

(i) Se existe discriminação de minorias e os fatores adequados foram controlados, qual é o sinal de β_1?

(ii) Faça a regressão de *approve* sobre branco e registre os resultados na forma usual. Interprete o coeficiente sobre *white*. Ele é estatisticamente significativo? Ele é grande do ponto de vista prático?

(iii) Como controles, adicione as variáveis *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2* e *vr*. O que acontece com o coeficiente de branco (*white*)? Ainda existem evidências de discriminação contra não brancos?

(iv) Agora, permita que o efeito de raça interaja com a variável que mede outras obrigações como uma porcentagem da renda (*obrat*). O termo de interação é significativo?

(v) Usando o modelo do item(iv), qual é o efeito de ser branco sobre a probabilidade de aprovação quando *obrat* = 32, que é praticamente o valor médio da amostra? Obtenha um intervalo de confiança de 95% para esse efeito.

C9 A questão da existência dos planos de pensão 401(k), disponíveis a muitos trabalhadores norte-americanos, de aumentar ou não as poupanças líquidas tem despertado muito interesse. O conjunto de dados 401KSUB contém informações sobre ativos financeiros líquidos (*nettfa*), renda familiar (*inc*), uma variável binária para elegibilidade em um plano 401(k) (*e401k*) e várias outras variáveis.

(i) Que proporção das famílias da amostra é elegível para participação em um plano 401(k)?

(ii) Estime um modelo de probabilidade linear explicando a elegibilidade ao 401(k) em termos de renda, idade e gênero. Inclua renda e idade em forma quadrática e registre os resultados na forma usual.

(iii) Você diria que a elegibilidade em um plano 401(k) é independente de renda e idade? E de gênero? Explique.

(iv) Obtenha os valores ajustados do modelo de probabilidade linear estimado no item (ii). Alguns dos valores ajustados são negativos ou maiores do que um?

(v) Usando os valores ajustados de $\widehat{e401k}_i$ do item (iv), defina $\widehat{e401k}_i = 1$ se $\widehat{e401k}_i \geq 0{,}5$, e $\widehat{e401k}_i = 0$ se $\widehat{e401k}_i < 0{,}5$. Das 9.275 famílias, quantas podemos prever como elegíveis para um plano 401(k)?

(vi) Das 5.638 famílias não elegíveis para um plano 401(k), qual porcentagem podemos prever que não terá um 401(k), usando o previsor $\widehat{e401k}_i$? Das 3.637 famílias elegíveis, qual é a porcentagem prevista que obterá um plano? (Se o seu programa econométrico tiver um comando "tabular" pode ser útil.)

(vii) A porcentagem geral corretamente prevista é de cerca de 64,9%. Você acha que é uma descrição completa de quão bem o modelo funciona, dadas as suas respostas do item (vi)?

(viii) Adicione a variável *pira* como uma variável explicativa ao modelo de probabilidade linear. Mantendo outros fatores iguais, se uma família tiver alguém com uma conta individual de aposentadoria, quão alta é a probabilidade estimada de essa família ser elegível para um plano 401(k)? O valor é estatisticamente diferente de zero a um nível de 10%?

C10 Use os dados do arquivo NBASAL neste exercício.

(i) Estime um modelo de regressão linear relacionando pontos por jogo com experiência na liga e posição (*guard*, *forward* ou *center*). Inclua a experiência em forma quadrática e use os centrais como grupo de base . Registre os resultados na forma usual.

(ii) Por que você não incluiu todas as três variáveis *dummy* de posição no item (i)?

(iii) Mantendo experiência fixa, um defensor pontua mais do que um central? Quanto a mais? A diferença é estatisticamente significativa?

(iv) Agora, adicione o estado civil à equação. Mantendo posição e experiência fixas, os jogadores casados são mais produtivos (com base nos pontos por jogo)?

(v) Adicione interações do estado civil com ambas as variáveis de experiência. Neste modelo expandido, existem evidências fortes de que o estado civil afeta os pontos por jogo?

(vi) Estime o modelo do item (iv), mas use assistências por jogo como variável dependente. Surgiram diferenças notáveis em relação ao item (iv)? Discuta.

C11 Use os dados do arquivo 401KSUBS para este exercício.

(i) Calcule a média, o desvio padrão, os valores mínimo e máximo de *nettfa* na amostra.

(ii) Teste a hipótese de que *nettfa* médio não difere por status de elegibilidade em planos 401(k); use uma alternativa bilateral. Qual é o montante em dólar da diferença estimada?

(iii) No item (ii) do Exercício em computador C9 fica claro que *e401k* não é exógena em um modelo de regressão simples; no mínimo, ela varia por renda e idade. Estime um modelo de regressão linear múltiplo para *nettfa* que inclua renda, idade e *e401k* como variáveis explicativas. As variáveis de renda e idade devem aparecer como quadráticas. Agora, qual é o efeito estimado em dólares da elegibilidade nos planos 401(k)?

(iv) Ao modelo estimado no item (iii), adicione as interações $e401k \cdot (age - 41)$ e $e401k \cdot (age - 41)^2$. Note que a idade média na amostra é cerca de 41, assim, no novo modelo, o coeficiente sobre *e401k* é o efeito estimado da elegibilidade na idade média. Qual termo de interação é significativo?

(v) Comparando as estimativas dos itens (iii) e (iv), os efeitos estimados da elegibilidade aos planos 401(k) na idade de 41 anos diferem muito? Explique.

(vi) Agora, retire os termos de interação do modelo, mas defina cinco variáveis dummy para o tamanho das famílias: *fsize1*, *fsize2*, *fsize3*, *fsize4* e *fsize5*. A variável *fsize5* é única para famílias com cinco ou mais membros. Inclua as *dummies* de tamanho de família no modelo estimado no item (iii); certifique-se de escolher um grupo-base. As variáveis *dummy* familiares são significativas a um nível de 1%?

(vii) Agora, faça um teste de Chow para o modelo

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 e401k + u$$

nas cinco categorias de tamanho de família, permitindo diferenças de intercepto. A soma dos quadrados dos resíduos restrita, SQR_r, é obtida pelo item (vi) porque aquela regressão supõe que todas as inclinações sejam as mesmas. A soma dos quadrados dos resíduos irrestrita é $SQR_{ir} = SQR_1 + SQR_2 + ... + SQR_5$, onde SQR_f é a soma dos quadrados dos resíduos da equação estimada usando somente o tamanho de família *f*. Você deve se convencer de que existem 30 parâmetros no modelo irrestrito (5 interceptos e 25 inclinações) e 10 parâmetros no modelo restrito (5 interceptos e 5 inclinações). Assim, o número de restrições testadas é $q = 20$, e o *gl* para o modelo irrestrito é $9.275 - 30 = 9.245$.

C12 Use os dados do arquivo BEAUTY, que contém um subconjunto de variáveis (mas mais observações aproveitáveis do que nas regressões) registrado por Hamermesh e Biddle (1994).

(i) Encontre as frações separadas de homens e mulheres que são classificadas com aparência acima da média. Existem mais pessoas classificadas com aparência acima ou abaixo da média?

(ii) Teste a hipótese nula de que as frações de população de mulheres e homens com aparência acima da média são as mesmas. Registre o *p*-valor unilateral de que a fração é maior para mulheres. (Dica: Estimar um modelo de probabilidade linear simples é mais fácil.)

(iii) Agora estime o modelo

$$\log(wage)\ \beta_0 + \beta_1 belavg + \beta_2 abvavg + u$$

separadamente para homens e mulheres e registre os resultados na forma usual. Em ambos os casos, interprete o coeficiente sobre *belavg*. Explique em palavras o que a hipótese H_0: $\beta_1 = 0$ contra H_1: $\beta_1 < 0$ quer dizer, e encontre os *p*-valores para homens e mulheres.

(iv) Existe alguma evidência convincente de que mulheres com aparência acima da média ganham mais do que mulheres com aparência média? Explique.

(v) Para homens e mulheres, adicione as variáveis explicativas *educ*, *exper*, $exper^2$, *union*, *goodhlth*, *black*, *married*, *south*, *bigcity*, *smllcity* e *service*. Os efeitos das variáveis de "aparência" mudam de forma importante?

(vi) Use a forma de SQR da estatística *F* de Chow para testar se as inclinações das funções de regressão do item (v) diferem entre homens e mulheres. Certifique-se de permitir um deslocamento de intercepto sob a hipótese nula.

C13 Use os dados do arquivo APPLE para responder a essa questão.

(i) Defina uma variável binária como *ecobuy* = 1, se *ecolbs* > 0, e *ecobuy* = 0, se *ecolbs* = 0. Em outras palavras, *ecobuy* indica se, a determinados preços, uma família compraria maçãs ecológicas. Que fração de famílias afirma que compraria maçãs com selos ecológicos?

(ii) Estime o modelo de probabilidade linear

$$ecobuy = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc + \beta_4 hhsize + \beta_5 educ + \beta_6 age + u,$$

e registre os resultados na forma usual. Interprete cuidadosamente os coeficientes sobre as variáveis de preço.

(iii) As variáveis que não são de preço são conjuntamente significativas no MPL? (Use a estatística *F* comum, mesmo que não seja válida quando há heteroscedasticidade.) Qual variável explicativa, além das variáveis de preço, parece ter o efeito mais importante sobre a decisão de comprar maçãs ecológicas? Isso faz sentido para você?

(iv) No modelo do item (ii), substitua *faminc* por log(*faminc*). Qual modelo adapta melhor esses dados, o que usa *faminc* ou log(*faminc*)? Interprete o coeficiente de log(*faminc*).

(v) Na estimação do item (iv), quantas probabilidades estimadas são negativas? Quantas são maiores do que um? Você deveria se preocupar?

(vi) Na estimação do item (iv), calcule o percentual corretamente previsto para cada resultado, *ecobuy* = 0 e *ecobuy* = 1. Qual resultado é mais bem previsto pelo modelo?

C14 Use os dados do arquivo CHARITY para responder a essa questão. A variável *respond* é uma variável *dummy* igual a um se a pessoa respondeu com uma contribuição à correspondência mais recente enviada por uma organização de caridade. A variável *resplast* é uma variável *dummy* igual a um se a pessoa respondeu ao envio anterior. *avggift* é a média de doações passadas (em florins holandeses), e *propresp* é a proporção de vezes que a pessoa respondeu às correspondências anteriores.

(i) Estime um modelo de probabilidade linear relacionando *respond* a *resplast* e *avggift*. Registre os resultados na forma usual e interprete o coeficiente sobre *resplast*.

(ii) O valor médio de doações passadas parece afetar a probabilidade de resposta?

(iii) Adicione a variável *propresp* ao modelo e interprete seu coeficiente. (Tenha cuidado aqui: o aumento de um em *propresp* é a maior mudança possível.)

(iv) O que aconteceu com o coeficiente de *resplast* quando *propresp* foi adicionada à regressão? Isso faz sentido?

(v) Adicione *mailsyear*, o número de correspondências enviadas por ano, ao modelo. Quão grande é seu efeito estimado? Por que essa pode não ser uma boa estimativa do efeito causal dos envios sobre as respostas?

C15 Use os dados do arquivo FERTIL2 para responder a essa questão.

(i) Encontre os menores e os maiores valores de *children* na amostra. Qual é a média de *children*? Alguma mulher tem exatamente o número médio de *children*?

(ii) Qual porcentagem de mulheres tem eletricidade em casa?

(iii) Calcule a média de *children* para aquelas que não têm eletricidade em casa e faça o mesmo com as que têm. Comente sobre seus achados. Teste se as médias da população são as mesmas usando uma regressão simples.

(iv) Usando o item (iii), você pode deduzir que ter eletricidade "faz com que" mulheres tenham menos filhos? Explique.

(v) Estime um modelo de regressão múltipla do tipo registrado na equação (7.37), mas adicione age^2, *urban* e as três *dummies* religiosas. De que forma o efeito estimado de ter eletricidade se compara com o do item (iii)? Ele ainda é estatisticamente significativo?

(vi) À equação do item (v), adicione uma interação entre *electric* e *educ*. Seu coeficiente é estatisticamente significativo? O que acontece com o coeficiente em eletric?

(vii) O valor mediano e a moda para *educ* é 7. Na equação do item (vi), use o termo de interação centrado $electric \cdot (educ - 7)$ no lugar de $electric \cdot educ$. O que acontece com o coeficiente em *electric* comparado com o do item (vi)? Por quê? De que forma o coeficiente sobre *electric* se compara com o do item (v)?

C16 Use os dados do arquivo CATHOLIC para responder a essa questão.

(i) Na amostra total, qual porcentagem de estudantes frequenta uma escola católica no ensino médio? Qual é a média de *math*12 na amostra total?

(ii) Faça uma regressão simples de *math*12 sobre *cathhs* e registre os resultados na forma usual. Interprete suas descobertas.

(iii) Agora, adicione as variáveis *lfaminc*, *motheduc* e *fatheduc* à regressão do item (ii). Quantas observações são usadas na regressão? O que acontece com o coeficiente de *cathhs*, junto com sua significância estatística?

(iv) Retorne à regressão simples de *math*12 sobre *cathhs*, mas restrinja as observações àquelas usadas na regressão múltipla do item (iii). Alguma conclusão importante muda?

(v) À regressão múltipla do item (iii), adicione interações entre *cathhs* e cada uma das outras variáveis explicativas. Os termos de interação são individual ou conjuntamente significativos?

(vi) O que acontece com o coeficiente sobre *cathhs* na regressão do item (v)? Explique por que esse coeficiente não é muito interessante.

(vii) Calcule o efeito parcial médio de *cathhs* no modelo estimado no item (v). Como isso se compara com os coeficientes de *cathhs* dos itens (iii) e (v)?