EDUARDO HORTA

ECONOMETRIA

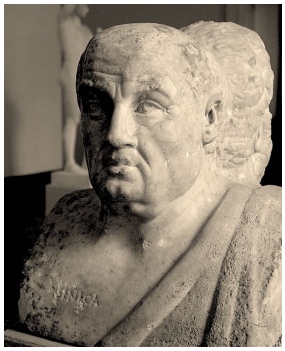
PPGE/UFRGS

PUBLISHED BY PPGE/UFRGS

First printing, October 2020

Sumário

The time will come when diligent research over long periods will bring to light things which now lie hidden. A single lifetime, even though entirely devoted to the sky, would not be enough for the investigation of so vast a subject... And so this knowledge will be unfolded only through long successive ages. There will come a time when our descendants will be amazed that we did not know things that are so plain to them... Many discoveries are reserved for ages still to come, when memory of us will have been effaced.



Lucius Annaeus Seneca, *Naturales quaestiones.*

Prelúdio

Um texto sobre Econometria — tanto mais um texto para alunos de pós-graduação — não deveria se esquivar de abordar questões epistemológicas: não resta dúvida de que esse é um assunto que desperta o interesse do estudante dotado de uma mente questionadora. Nosso tempo é exíguo, contudo, e me resta indicar aos curiosos um ponto de partida para a literatura relevante.[1] Um texto sobre Econometria tampouco deveria se esquivar de apresentar um panorama da pesquisa — teórica e aplicada — nessa disciplina. Com o risco de levar a pecha de pedante — afinal, para que mencionar aquilo que se deve fazer e não fazê-lo? — recorro novamente ao expediente de apenas indicar uma referência importante, qual seja, o *review* escrito por John Geweke, Joel Horowitz e Hashem Pesaran em 2006 com o título *Econometrics: A Bird's Eye View* e republicado em 2008 no *New Palgrave Dictionary of Economics*[2] com um título um tanto mais prosaico.

O leitor já pode deduzir, após o parágrafo acima, que esse texto apresentará a teoria Econométrica desde a perspectiva do formalismo matemático — é isso que o tempo nos permite, e é a partir do entendimento desse formalismo que se pode efetivamente compreender questões relativas à teoria do conhecimento e à metodologia científica, bem como proceder a aplicações que não se restrinjam a mero mimetismo ("$p < 0.01 \rightarrow$ posso rejeitar a hipótese nula!"). Ainda assim, há pelo menos uma questão de caráter epistemológico sobre a qual eu eu gostaria de comentar antes de prosseguir, que é a seguinte: no jargão econométrico, o conceito de *Data Generating Process* (DGP) é amplamente empregado, em que pese o fato de ser uma noção imprecisa. Todavia, em que pese o fato de ser uma noção

[1] Trygve Haavelmo. The Probability Approach in Econometrics. *Econometrica*, 12:iii–115, 1944; Olav Bjerkholt. On the Founding of the Econometric Society. *Journal of the History of Economic Thought*, 39(2), 2017; David F. Hendry. Econometrics – Alchemy or Science? *Economica*, 47 (188):387–406, 1980; and Kevin D. Hoover. The Methodology of Econometrics. In Terence C. Mills and Kerry Patterson, editors, *Palgrave Handbook of Econometrics*, volume 1. Palgrave Macmillan UK, 2006

[2] John Geweke, Joel L. Horowitz, and Hashem Pesaran. *Econometrics*, pages 1–44. Palgrave Macmillan UK, London, 2017

imprecisa, devemos admitir que há muito didatismo em se considerar que a realidade/natureza/Τύχη (Fortuna) gera os fenômenos que observamos utilizando um mecanismo capaz de produzir observáveis que *parecem* aleatórios (e.g., um programa de computador), mecanismo esse configurado com parâmetros por ela escolhidos (e, para nós, desconhecidos), e que nosso objetivo nesse jogo é oferecer um *educated guess* sobre qual o valor desses parâmetros, baseado nas observações (por Τύχη reveladas) a que temos acesso. Todavia, há pelo menos dois inconvenientes nessa ideia que merecem ser mencionados, ambos relacionados à dificuldade de se estabelecer uma ponte entre um discurso sobre fenômenos de um lado e um discurso delimitado por uma linguagem formal de outro:

1. ***Quantidades numéricas observadas no mundo real não são variáveis aleatórias***! Variáveis aleatórias[3] são objetos constritos a uma linguagem formal/axiomática. *Por exemplo*, a *'taxa de juros de amanhã'* é um valor numérico *in potentia*, sobre o qual temos incerteza, mas não faz sentido perguntar se é uma função mensurável definida em um espaço de probabilidade, etc. Essa distinção costumeiramente é deixada de lado, e dizemos despreocupadamente que a *'taxa de juros de amanhã'* **é** uma variável aleatória. É importante ressaltar que é imaterial, no exemplo acima, o fato de estarmos considerando um evento futuro; poderíamos com igual valor didático tomar a *'taxa de juros de ontem'* se, por alguma razão, desconhecêssemos essa quantidade.

2. Sentenças da forma "A probabilidade de E é p", onde E é um evento exprimível em linguagem natural e p é um número real entre 0 e 1, ***tipicamente não fazem referência a um atributo do mundo real***. *Por exemplo*, considere o experimento que consiste em lançar uma moeda equilibrada duas vezes consecutivas, em lançamentos fisicamente independentes. Um modelo formal para esse experimento é o seguinte: consideramos o espaço amostral $\Omega = \{HH, HT, TH, TT\}$ (H para *cara*, T para *coroa*) munido da medida de probabilidade $\mathbb{P}$ definida via $\mathbb{P}\{\omega\} = 1/4$ para $\omega \in \Omega$. Nesse contexto, o evento (expresso em linguagem natural) "observar *cara* no primeiro lançamento" corresponde ao evento $\{HH, HT\}$, e a sentença "a probabilidade de observar-

Τύχη de Antioquia, cópia romana de um original em bronze feito por Eutiquides, *Galleria dei Candelabri*.

[3] Lembre que uma variável aleatória é uma função $x\colon \Omega \to \mathbb{R}$, onde Ω é um conjunto não-vazio munido de uma σ-álgebra $\mathscr{A}$, tal que os conjuntos

$$[x \leq \xi] \coloneqq \{\omega \in \Omega\colon\ x(\omega) \leq \xi\}$$

são **eventos** (isto é, elementos de $\mathscr{A}$), para qualquer $\xi \in \mathbb{R}$. Ao conjunto Ω chamamos de **espaço amostral** e consideramos que ele representa a coleção de todos os possíveis *estados do mundo*.

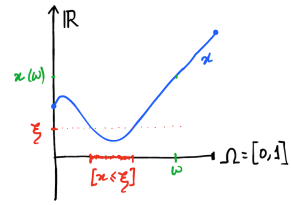


Figura 1: Exemplo de uma variável aleatória no espaço amostral $\Omega = [0, 1]$.

mos *cara* no primeiro lançamento é (igual a) 1/2" corresponde à sentença formal

$$\text{“}\mathbb{P}\{HH, HT\} = 1/2\text{”}.$$

Note, porém que a função $\mathbb{P}$ acima foi *escolhida*, dentro do escopo de uma linguagem formal, para satisfazer nossa intuição de qual seria um modelo adequado para descrever o fenômeno sob consideração, de acordo com aquilo que o enunciado do problema deixa transparecer: "moeda equilibrada", "lançamentos independentes", etc. Note também que, se formos exigentes no nosso uso da linguagem, deveríamos traduzir a sentença "$\mathbb{P}\{HH, HT\} = 1/2$" como "a $\mathbb{P}$-probabilidade de observarmos *cara* no primeiro lançamento é (igual a) 1/2". Esse exemplo deixa transparecer — em que pese o fato de que o modelo escolhido satisfaz, sim, um critério (informal) de razoabilidade — que o valor numérico 1/2 aparecendo na sentença acima não é "algo que está no mundo". Sendo assim, o melhor é interpretar esse valor numérico como uma quantificação de nossa incerteza a respeito de um fenômeno que temos interesse em modelar, e essa quantificação depende de uma escolha (nossa!) de qual modelo formal vamos utilizar: em suma, sentenças de probabilidade (*probability statements*) são epistêmicas! Em suma, sentenças do tipo "o verdadeiro valor do parâmetro é p" nada mais são, na maioria das vezes, do que uma abstração, referentes a objetos que existem apenas dentro do escopo de uma linguagem formal.[4] Como disse o célebre George Box,[5]

"All models are wrong, but some are useful."

É importante ter isso em mente quando falamos, por exemplo, sobre "o verdadeiro valor de β", etc.

[4] Embora, no caso de populações finitas, parâmetros teóricos correspondam a quantidades existentes na realidade.

[5] George Box. Robustness in the Strategy of Scientific Model Building. In Robert L. Launer and Graham N. Wilkinson, editor, *Robustness in Statistics*, pages 201 – 236. Academic Press, 1979

Sobre notações e outras noções

Após algum tempo convivendo com notações matemáticas é inevitável dar-se conta de que, como a busca pelo Santo Graal, o desejo por uma notação abrangente e unificada não é mais que uma busca vã. Em Probabilidade Axiomática, a notação canônica (se é que existe tal coisa) recomenda que variáveis aleatórias sejam denotadas por letras romanas maiúsculas, e.g. X, Y, Z; que valores realizados

(números reais) sejam denotados por letras romanas minúsculas, e.g. x, y, z etc. Na literatura de Econometria, todavia, consagrou-se a heresia em que variáveis aleatórias são denotadas por letras minúsculas, e nesse texto seguiremos essa tradição: os símbolos x, y, z, w, u, v (possivelmente acrescidos de um subíndice i ou ij) aqui denotam variáveis aleatórias; se estiverem em negrito (e.g., $\boldsymbol{x}$) representam *vetores* aleatórios; imitando o muito didático Prof. David Williams,[6] usaremos os símbolos $x^{\text{obs}}, y^{\text{obs}}, z^{\text{obs}}, w^{\text{obs}}, u^{\text{obs}}, v^{\text{obs}}$ (possivelmente acrescidos de um subíndice i ou ij) para denotar valores realizados[7] das variáveis aleatórias x, y, z, w, u, v, respectivamente; os símbolos ξ, η, ζ também denotam números reais genéricos, variáveis de integração etc (com sorte não precisaremos muito deles); os símbolos i, j, k, ℓ, n, m, d denotam números inteiros, índices etc.[8] Nisso, orbitaremos — mas não excessivamente de perto — a notação utilizada por Stachurski[9] e Hayashi.[10] Em particular, dado um **tamanho de amostra** $n \in \mathbb{N}$, e dadas variáveis aleatórias $y_1, \ldots, y_n$ e vetores aleatórios $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, onde

$$\boldsymbol{x}_i' = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{id} \end{pmatrix}, \quad i \in \{1, \ldots, n\},$$

escreveremos

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{e} \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

onde $\boldsymbol{x}'$ denota o transposto do vetor $\boldsymbol{x}$ (e, semelhantemente, $\boldsymbol{X}'$ denota a transposta da matriz $\boldsymbol{X}$). Note que $\boldsymbol{y}$ é um vetor $n \times 1$, enquanto $\boldsymbol{x}_i$ é um vetor $d \times 1$. Note também que o correto aqui seria escrever $x_{i,j}$ ao invés de x_{ij}, mas nos permitiremos esse ligeiro abuso de notação onde não houver possibilidade de trocarmos os pés pelas mãos (por exemplo, de confundirmos o número 11 com o par $1, 1$).

UM CONCEITO FUNDAMENTAL na literatura de modelos de regressão, a *esperança condicional* é definida da seguinte maneira: seja y uma variável aleatória **integrável** (quer dizer, y é tal que $\mathbb{E}(|y|) < \infty$), e seja $\boldsymbol{x}$ um vetor aleatório com valores em $\mathbb{R}^d$. A **es-**

[6] David Williams. *Weighing the Odds: A Course in Probability and Statistics*. Cambridge University Press, 2001

[7] Na prática, o conceito de "valor realizado" é vago; formalmente, x^{obs}, y^{obs}, z^{obs}, w^{obs}, u^{obs}, v^{obs} denotam números reais.

[8] Comparativamente, enquanto na notação canônica escreveríamos, por exemplo, $F_Y(y) = \mathbb{P}(Y \leq y)$ para denotar o valor da função de distribuição acumulada F_Y da variável aleatória Y no ponto $y \in \mathbb{R}$ (respectivamente, $\mathbb{E}(X) = \int_{-\infty}^{+\infty} x\, f_X(x)\, \mathrm{d}x$ para denotar o valor esperado da variável aleatória absolutamente contínua X em termos de sua função densidade de probabilidade f_X), com a notação que seguiremos nessas notas escreveremos $F_y(\xi) = \mathbb{P}(y \leq \xi)$ para denotar o valor da função de distribuição acumulada F_y da variável aleatória y no ponto $\xi \in \mathbb{R}$)(respectivamente, $\mathbb{E}(x) = \int_{-\infty}^{+\infty} \xi\, f_x(\xi)\, \mathrm{d}\xi$ para denotar o valor esperado da variável aleatória x em termos de sua função densidade de probabilidade f_x.)

[9] John Stachurski. *A primer in econometric theory*. MIT Press, 2016

[10] Fumio Hayashi. *Econometrics*. Princeton University Press, 2000

perança condicional de y dado $\boldsymbol{x}$ é a variável aleatória $\mathbb{E}(y \,|\, \boldsymbol{x})$ que satisfaz as seguintes condições:

EC1. Tem-se que $\mathbb{E}(y \,|\, \boldsymbol{x}) = \varphi \circ \boldsymbol{x}$ para alguma função mensurável φ definida em $\mathbb{R}^d$ e tomando valores na reta, com $\mathbb{E}(|\varphi \circ \boldsymbol{x}|) < \infty$.

EC2. Vale a igualdade $\mathbb{E}\big(\mathbb{E}(y \,|\, \boldsymbol{x})\, \mathbb{I}_{[\boldsymbol{x} \in B]}\big) = \mathbb{E}(y\, \mathbb{I}_{[\boldsymbol{x} \in B]})$ para todo subconjunto Boreliano $B \subseteq \mathbb{R}^d$.

A função φ dada acima[11] é chamada a **função de regressão de y em $\boldsymbol{x}$**. Nessas condições, se $\boldsymbol{x}^{\text{obs}} \in \mathbb{R}^d$ é um possível valor assumido pelo vetor aleatório $\boldsymbol{x}$, definimos

$$\mathbb{E}(y \,|\, \boldsymbol{x} = \boldsymbol{x}^{\text{obs}}) \coloneqq \varphi(\boldsymbol{x}^{\text{obs}}).$$

Atenção para não se confundir!

Remark 1. Por definição, dadas funções $f\colon A \to B$ e $h\colon B \to C$, denotamos por $h \circ f$ a função cujo domínio é A, cujo contradomínio é C e cujo valor no ponto $a \in A$ é o elemento $h(f(a)) \in C$; isto é, $h \circ f(a) \coloneqq h(f(a))$. É costumeiro, em Probabilidade, Estatística, Econometria etc, escrever $g(\boldsymbol{x})$ ao invés de $g \circ \boldsymbol{x}$ quando $\boldsymbol{x}$ é um vetor aleatório de dimensão d e g é uma função de $\mathbb{R}^d$ em $\mathbb{R}$. Isso é um abuso de notação. Aqui, a fim de evitar ambiguidades, escreveremos $g \circ \boldsymbol{x}$ sempre que possível.

PARA FINALIZAR, será conveniente fazer a seguinte suposição ao longo de todo o texto:

> TODAS AS VARIÁVEIS ALEATÓRIAS QUE CONSIDERAREMOS TÊM VARIÂNCIA FINITA.[12]

Digo que essa suposição é conveniente por duas razões: primeiro, porque ela implica integrabilidade das variáveis aleatórias que consideraremos (e assim estaremos poupados, ao percorrer o texto, de *exigir* integrabilidade de y antes de nos permitirmos escrever $\mathbb{E}(y)$, por exemplo); segundo, porque ela nos permite[13] escrever $\mathbb{V}(\boldsymbol{x})$ para denotar a **matriz de covariância** de um vetor aleatório $\boldsymbol{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_d \end{pmatrix}'$:

$$\mathbb{V}(\boldsymbol{x}) \coloneqq \begin{pmatrix} \mathbb{V}(x_1) & \operatorname{cov}(x_1, x_2) & \cdots & \operatorname{cov}(x_1, x_d) \\ \operatorname{cov}(x_2, x_1) & \mathbb{V}(x_2) & \cdots & \operatorname{cov}(x_2, x_d) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(x_d, x_1) & \operatorname{cov}(x_d, x_2) & \cdots & \mathbb{V}(x_d) \end{pmatrix}$$

[11] É bem sabido que tal φ é essencialmente única, no seguinte sentido: se $g\colon \mathbb{R}^d \to \mathbb{R}$ é outra função (mensurável) satisfazendo a condição

$$\mathbb{E}\{g \circ \boldsymbol{x}\, \mathbb{I}_{[\boldsymbol{x} \in B]}\} = \mathbb{E}(y\, \mathbb{I}_{[\boldsymbol{x} \in B]})$$

para todo Boreliano $B \subseteq \mathbb{R}^d$, então vale que

$$\mathbb{P}(g \circ \boldsymbol{x} = \varphi \circ \boldsymbol{x}) = 1,$$

onde $\mathbb{I}_{[\boldsymbol{x} \in B]}$ é a variável aleatória definida pondo-se $\mathbb{I}_{[\boldsymbol{x} \in B]}(\omega) = 1$ se $\boldsymbol{x}(\omega) \in B$ e $\mathbb{I}_{[\boldsymbol{x} \in B]}(\omega) = 0$ se $\boldsymbol{x}(\omega) \notin B$.

[12] Define-se a **variância** de uma variável aleatória y como sendo a quantidade $\mathbb{V}(y)$ dada por

$$\mathbb{V}(y) \coloneqq \mathbb{E}(y^2) - (\mathbb{E}y)^2.$$

Essa quantidade é não-negativa mas, possivelmente, infinita.

[13] De fato, a desigualdade de Cauchy–Schwarz nos dá

$$|\operatorname{cov}(x, y)|^2 \leq \mathbb{V}(x)\mathbb{V}(y).$$

onde $\operatorname{cov}(x, y) \coloneqq \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y)$ para variáveis aleatórias x e y. Escreveremos, ainda,

$$\operatorname{cov}(\boldsymbol{x}, \boldsymbol{u}) \coloneqq \mathbb{E}(\boldsymbol{x}\boldsymbol{u}') - \mathbb{E}(\boldsymbol{x})\mathbb{E}(\boldsymbol{u}'),$$

quando $\boldsymbol{x}$ e $\boldsymbol{u}$ forem vetores aleatórios. No caso em que $\boldsymbol{u}$ tem dimensão 1 — digamos, $\boldsymbol{u} = (u)$ — escrevemos simplesmente $\operatorname{cov}(\boldsymbol{x}, u)$ e desconsideramos os sinais de transposição na definição acima.

Exercícios

Quaisquer erros nos enunciados são — obviamente — intencionais.

Exercício 1. Seja $\Omega = (0, 1]$ munido de uma σ-álgebra $\mathscr{A}$ tal que todo sub-intervalo $(a, b] \subseteq \Omega$, com $0 < a \leq b \leq 1$, satisfaz a relação $(a, b] \in \mathscr{A}$. Verifique que a função $x\colon \Omega \to \mathbb{R}$ definida, para $\omega \in \Omega$, por $x(\omega) = \omega^2$ é uma variável aleatória (com respeito a $\mathscr{A}$).

Exercício 2. Sejam x e y variáveis aleatórias. Verifique que

$$\mathbb{E}\big[\big(x - \mathbb{E}(x)\big) \cdot \big(y - \mathbb{E}(y)\big)\big] = \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y)$$

Exercício 3. Considere os vetores aleatórios $\boldsymbol{y} = (y_1 \quad \cdots y_n)'$, $\boldsymbol{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_d \end{pmatrix}'$ e $\boldsymbol{x}_i = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{id} \end{pmatrix}'$, $i \in \{1, \ldots, n\}$, e seja $\boldsymbol{X}$ a matriz cuja componente ij é x_{ij}. Sejam ainda Σ uma matriz $m \times d$ e $\boldsymbol{\xi} \in \mathbb{R}^m$, ambos não aleatórios. Verifique as seguintes igualdades:

1. $\mathbb{E}(\Sigma \boldsymbol{X}) = \Sigma \mathbb{E}(\boldsymbol{X})$, onde — por definição — a esperança de uma matriz aleatória é tomada componente-a-componente.
2. $\mathbb{E}(\boldsymbol{X}') = \big(\mathbb{E}(\boldsymbol{X})\big)'$.
3. $\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \boldsymbol{X}'\boldsymbol{X}$.
4. $\sum_{i=1}^{n} \boldsymbol{x}_i y_i = \boldsymbol{X}'\boldsymbol{y}$.
5. $\mathbb{V}(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}') - \mathbb{E}(\boldsymbol{x})\mathbb{E}(\boldsymbol{x}')$.
6. $\mathbb{V}(\Sigma \boldsymbol{x} + \boldsymbol{\xi}) = \Sigma \mathbb{V}(\boldsymbol{x}) \Sigma'$.

O modelo de regressão linear

"... avistamos de longe e dizemos, O Templo, quando entramos no Pátio dos Gentios tornamos a dizer, O Templo, e agora o carpinteiro José, apoiado à balaustrada, olha e diz, O Templo, e é ele quem tem razão..."'

José Saramago, *O Evangelho Segundo Jesus Cristo*

Ao longo de todo o texto, consideramos fixado um espaço de probabilidade $(\Omega, \mathscr{A}, \mathbb{P})$ no qual estão definidas todas as variáveis aleatórias que encontraremos. Ademais, no que segue n é um número natural, $y, y_1, \ldots, y_n$ e $u, u_1, \ldots, u_n$ são variáveis aleatórias, $\boldsymbol{x}$, $\boldsymbol{x}_1$, $\ldots$, $\boldsymbol{x}_n$ são vetores aleatórios de dimensão $d+1$,[14] onde $d \in \mathbb{N}$ etc. Por conveniência, assumiremos que a primeira componente desses vetores aleatórios é sempre uma constante e, guiados por elevado senso estético, não poderíamos tomar essa constante como sendo qualquer outro valor que não o número 1. Assim temos, por exemplo,

$$\boldsymbol{x}' = \begin{pmatrix} 1 & x_1 & \ldots & x_d \end{pmatrix}, \quad \boldsymbol{x}_i' = \begin{pmatrix} 1 & x_{i1} & \ldots & x_{id} \end{pmatrix}$$

e assim por diante.

[14] Isso representa um pequeno ajuste com respeito à notação introduzida no Prelúdio.

Um caso simples, para começar

Pode-se dizer que lançar os olhos sobre um texto de Econometria é o mesmo que iniciar uma contagem regressiva cujo estopim se dá no momento em que o leitor se depara com a equação

$$y = \beta_0 + \beta_1 x_1 + u. \tag{1}$$

Comemorai! Já é finda a contagem! Aí está aquele que podemos chamar de "O Modelo de regressão linear" (ao menos em uma versão bem simplezinha).[15] Há diversas maneiras de se chegar a uma tal igualdade, dependendo da ordem com que os ingredientes envolvidos são invocados. Podemos, por exemplo:

[15] Havíamos de começar em algum lugar, e decidimos começar por uma "amostra de tamanho 1" (por isso dispensamos o índice i) e com apenas um regressor não constante.

1. obtê-la como uma ***definição***, considerando $\boldsymbol{x}$ e u como dados, β_0 e β_1 como constantes pré-fixadas e *definindo* y por $y \coloneqq \beta_0 + \beta_1 x_1 + u$.

2. atingi-la por uma ***proposição***, e.g. tomando $\boldsymbol{x}$ e y como dados e demonstrando que existem números reais β_0 e β_1 e uma variável aleatória u satisfazendo $\mathbb{E}(u) = 0$ e $\operatorname{cov}(u, x_1) = 0$, tal que $y = \beta_0 + \beta_1 x_1 + u$.

3. considerar que a igualdade (1) é uma ***suposição*** (uma relação hipotética entre essas variáveis); isso usualmente é feito em um contexto aplicado, onde $\boldsymbol{x}$ e y modelam quantidades do mundo real, e o pesquisador escreve "suponha que $y = \beta_0 + \beta_1 x_1 + u$ para algum par de parâmetros reais β_0 e β_1 e alguma variável aleatória u satisfazendo isso, isso e aquilo". É... o papel aceita tudo!

Não há muito mais o que dizer[16], a não ser que podemos dar o pontapé inicial no nosso plano de tudo denotar em notação matricial e reescrever (1) como

$$y = \boldsymbol{x}'\boldsymbol{\beta} + u. \tag{2}$$

[16] Claro, há o jargão: y é dito a **variável resposta** ou o **regressando**; $\boldsymbol{x}$ é dito o vetor de **covariáveis** / **preditores** / **regressores** etc; u é dito o **termo de erro**.

Agora sim: O Modelo! Daqui, temos dois resultados que ilustram dois dos itens que discutimos logo acima:

Teorema 2. Considere um vetor aleatório $\boldsymbol{x}' = \begin{pmatrix} 1 & x_1 \end{pmatrix}$ e uma variável aleatória u, e sejam β_0 e β_1 números reais arbitrários mas fixados. Seja ainda y a variável aleatória *definida* pela equação

$$y \coloneqq \beta_0 + \beta_1 x_1 + u.$$

Se $\mathbb{E}(u) = 0$ e $\operatorname{cov}(x_1, u) = 0$, então

$$\operatorname{cov}(x_1, y) = \beta_1 \mathbb{V}(x_1) \quad \text{e} \quad \mathbb{E}(y) = \beta_0 + \beta_1 \mathbb{E}(x_1).$$

Teorema 3. Considere um vetor aleatório $\boldsymbol{x}' = (1 \quad x_1)$ e uma variável aleatória y, e assuma que $\mathbb{V}(x_1) > 0$. Sejam ainda β_0 e β_1 *definidos* respectivamente por

$$\beta_1 \coloneqq \frac{\operatorname{cov}(x_1, y)}{\mathbb{V}(x_1)} \quad \text{e} \quad \beta_0 \coloneqq \mathbb{E}(y) - \beta_1 \mathbb{E}(x_1).$$

Então, a variável aleatória u *definida* por

$$u \coloneqq y - (\beta_0 + \beta_1 x_1)$$

satisfaz $\mathbb{E}(u) = 0$ e $\operatorname{cov}(x_1, u) = 0$.

O primeiro dos dois teoremas acima nos diz que, se supusermos que u e x_1 são não-correlacionados então, ao definir y como no enunciado, estamos imediatamente forçando a covariância entre x_1 e y a assumir um valor bem específico — a saber, $\beta_1 \mathbb{V}(x_1)$ — e, semelhantemente, forçamos $\mathbb{E}(y)$ a assumir um valor bem específico — a saber, $\beta_0 + \beta_1 \mathbb{E}(x_1)$. Reciprocamente, o Teorema 3 nos garante que, se y e $\boldsymbol{x}$ são quaisquer,[17] então sempre é possível representar y como uma transformação linear de $\boldsymbol{x}$ acrescida de um termo de erro com média zero e não-correlacionado com $\boldsymbol{x}$ — em certo sentido, isso é dizer que O Modelo de regressão linear está sempre corretamente especificado, e isso nos leva ao terceiro dos itens que discutimos anteriormente: a ironia de que "o papel aceita tudo" foi uma injustiça minha contra aqueles que assumem *coisas* em seus modelos, pois o Teorema 3 os autoriza a escrever a equação (1) sem apelar à necromancia ou às artes ocultas.[18] Agora, um exemplo.

[17] Não se esqueça da suposição que fizemos — EM VERMELHO — no Prelúdio a essas notas.

[18] Mais adiante, quando falarmos sobre endogeneidade e variáveis instrumentais, voltaremos a desconfiar que "o papel aceita tudo".

Exemplo 4. Esse poderia facilmente levar a pecha de ser um exemplo artificial, e a acusação não seria sem merecimento! Em qualquer caso, vamos prosseguir e esperar que algo se esclareça com ele. Considere duas variáveis aleatórias x_1 e v mutuamente independentes e com distribuições simétricas em torno de zero, e seja y a variável aleatória *definida* por

$$y \coloneqq x_1^2 + v.$$

Afirmo que $\operatorname{cov}(x_1, y) = 0$. De fato,

$$\begin{aligned} \operatorname{cov}(x_1, y) &= \operatorname{cov}(x_1, x_1^2 + v) \\ &= \operatorname{cov}(x_1, x_1^2) + \operatorname{cov}(x_1, v) \\ &= \mathbb{E}(x_1^3) = 0 \end{aligned}$$

Sendo assim, ao escrevermos $y = \beta_0 + \beta_1 x_1 + u$, com β_0, β_1 e u definidos como no Teorema 3, realmente não estamos fazendo muita coisa, pois aqui ocorre que $\beta_1 = 0$, $\beta_0 = \mathbb{E}(y)$, $u = y - \mathbb{E}(y)$ e aí (pasmem!) $y = \mathbb{E}(y) + y - \mathbb{E}(y)$.

COM ISSO, ENCERRAMOS a apresentação d'O Modelo em sua versão bem simplezinha — isso tudo foi apenas um aquecimento, para que o leitor pudesse se sentir seguro uma última vez antes de nos aventurarmos nos mares turbulentos da Notação Matricial.

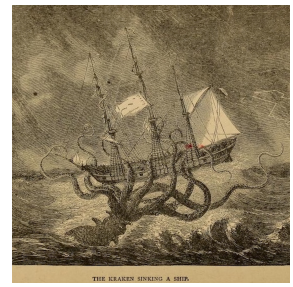


Figura 2: J.W. Buel, *Sea and Land*. Muitos alunos de Econometria relatam ter pesadelos em que a Notação Matricial aparece no lugar do Kraken.

Agora temos uma amostra!

Suponhamos agora que $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ satisfazem as igualdades

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \qquad i \in \{1, \ldots, n\}, \tag{3}$$

(O Modelo!) onde $u_1, u_2, \ldots, u_n$ são termos de erro e onde $\boldsymbol{\beta}' := (\beta_0 \quad \beta_1 \quad \cdots \quad \beta_d) \in \mathbb{R}^{d+1}$. Estaremos principalmente[19] interessados no caso em que os vetores aleatórios

$$(\boldsymbol{x}_1 \quad y_1)', \ldots, (\boldsymbol{x}_n \quad y_n)' \tag{4}$$

formam uma **amostra aleatória**, isto é, no caso em que sua distribuição conjunta é da forma

$$\begin{aligned}\mathbb{P}\left\{\bigcap_{i=1}^{n}[\boldsymbol{x}_i \in A_i, y_i \in B_i]\right\} &= \prod_{i=1}^{n}\mathbb{P}(\boldsymbol{x}_i \in A_i, y_i \in B_i) \\ &= \prod_{i=1}^{n}\mathbb{P}(\boldsymbol{x}_1 \in A_i, y_1 \in B_i),\end{aligned}$$

onde A_i e B_i, $i \in \{1, \ldots, n\}$ denotam subconjuntos Borelianos de $\mathbb{R}^{d+1}$ e de $\mathbb{R}$, respectivamente. A primeira das duas igualdades acima caracteriza a **independência** entre os vetores aleatórios envolvidos; a segunda nos diz que eles são **identicamente distribuídos**. Por essa razão, quando os vetores em (4) formam uma amostra aleatória, dizemos equivalentemente que eles são **iid**.[20]

No Prelúdio havíamos convencionado escrever (fazendo aqui uma

[19] Mas não exclusivamente!

[20] Fica a cargo do leitor decifrar a sigla.

ligeira correção)

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{e} \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} = \begin{pmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1,d} \\ x_{2,0} & x_{2,1} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nd} \end{pmatrix}$$

e, acrescentando agora ao nosso estoque mais uma notação, a saber,

$$\boldsymbol{u}' = \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix},$$

podemos finalmente — apoiados à balaustrada — dizer, O Modelo!, e contemplá-lo em toda sua generalidade:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{5}$$

O leitor familiarizado com a Álgebra Linear vai se recordar que há uma dualidade entre sistemas de equações lineares e produtos de matrizes por vetores. Algo parecido ocorre aqui, e é um exercício verificar que as equações (3) e (5) são de fato equivalentes.

Os Teoremas 2 e 3 acima abordam a questão de como obter O Modelo a depender da ordem com que são acrescidos os ingredientes envolvidos. Podemos seguir o mesmo roteiro para obter a equação (5), e é isso que faremos agora.

Teorema 5. Considere uma matriz aleatória $\boldsymbol{X}$ de dimensões $n \times (d+1)$ e um vetor aleatório $\boldsymbol{u}$ em $\mathbb{R}^n$, e seja $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ um vetor arbitrário mas fixado. Seja ainda $\boldsymbol{y}$ o vetor aleatório *definido* pela igualdade (5). Se $\mathbb{E}(\boldsymbol{u}) = \boldsymbol{0} \in \mathbb{R}^n$ e $\mathbb{E}(\boldsymbol{X'u}) = \boldsymbol{0} \in \mathbb{R}^{d+1}$, então

$$\mathbb{E}(\boldsymbol{y}) = \mathbb{E}(\boldsymbol{X})\boldsymbol{\beta} \quad \text{e} \quad \mathbb{E}(\boldsymbol{X'X})\boldsymbol{\beta} = \mathbb{E}(\boldsymbol{X'y}) \tag{6}$$

Teorema 6. Considere uma matriz aleatória $\boldsymbol{X}$ de dimensões $n \times (d+1)$ e um vetor aleatório $\boldsymbol{y}$ em $\mathbb{R}^n$, e assuma que $\mathbb{E}(\boldsymbol{X'X})$ é invertível. Seja ainda $\boldsymbol{\beta}$ *definido* por

$$\boldsymbol{\beta} = \big(\mathbb{E}(\boldsymbol{X'X})\big)^{-1}\mathbb{E}(\boldsymbol{X'y}). \tag{7}$$

Então o vetor aleatório $\boldsymbol{u}$ *definido* por

$$\boldsymbol{u} \coloneqq \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$$

satisfaz $\mathbb{E}(\boldsymbol{X'u}) = \mathbf{0} \in \mathbb{R}^{d+1}$. Adicionalmente, se

$$\boldsymbol{x}_i' = (1 \quad x_{i1} \quad \cdots \quad x_{id})$$

para todo i (isto é, a primeira coluna de $\boldsymbol{X}$ é uma coluna de 1's), e se os vetores em (4) são identicamente distribuídos, então $\mathbb{E}(\boldsymbol{u}) = \mathbf{0} \in \mathbb{R}^n$.

O Teorema 6 diz que, até certo ponto, O Modelo linear com regressores exógenos[21] está sempre "corretamente especificado", no sentido de que exprime uma relação válida entre $\boldsymbol{y}$ e $\boldsymbol{x}$; de fato, nos diz explicitamente quem devem ser $\boldsymbol{\beta}$ e $\boldsymbol{u}$ para que tal relação valha. Uma outra maneira de se enxergar esse fato é a seguinte:

[21] Mais sobre jargões logo adiante!

Corolário 7. Considere uma matriz aleatória $\boldsymbol{X}$ de dimensões $n \times (d+1)$ e um vetor aleatório $\boldsymbol{y}$ em $\mathbb{R}^n$, e assuma que $\mathbb{E}(\boldsymbol{X'X})$ é invertível. Então existe um "vetor de parâmetros" $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ e termos de erro $u_1, \ldots, u_n$ tais que $\mathbb{E}(\boldsymbol{X'u}) = 0$ e $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$.

Até aqui, fizemos considerações "teóricas", no sentido de que estamos calculando esperanças, covariâncias etc. Isso nada tem a ver, ainda, com o método de mínimos quadrados ordinários — o qual, em particular, pode ser apresentado sem que seja feita qualquer menção a probabilidades, variáveis aleatórias e *tutti quanti*: de fato, MQO pode ser visto meramente como um método numérico para encontrar um hiperplano afim que se "ajusta de maneira ótima" a um conjunto de pontos no espaço Euclidiano. A despeito disso, a riqueza do método se revela em toda a sua magnitude precisamente quando o enxergamos sob um ponto de vista probabilístico, e é isso que faremos a seguir.

Exercícios

Quaisquer erros nos enunciados são — evidentemente — intencionais.

Exercício 4. Demonstre o Teorema 2.

Exercício 5. Demonstre o Teorema 3.

Exercício 6. Verifique que as equações (3) e (5) são equivalentes.

Exercício 7. Demonstre o Teorema 5

Exercício 8. Mostre[22] que, se a suposição na segunda parte do Teorema 6 for satisfeita (isto é, o vetor de regressores inclui uma constante e os dados são identicamente distribuídos), então

$$\boldsymbol{\beta} = \left(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\right)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1).$$

Ademais, utilize a igualdade acima para resolver explicitamente a expressão para $\boldsymbol{\beta}$ no caso particular em que $d = 1$ — isto é, no caso em que $\boldsymbol{x}_1' = (1 \quad x_{1,1})$.

[22] Dica: use o Exercício 3.

Exercício 9. Demonstre o Teorema 6.

Exercício 10. Seja $\boldsymbol{x}$ um vetor aleatório de dimensão d.

1. Mostre que $\mathbb{V}(\boldsymbol{x})$ é positiva semidefinida (isto é, a forma quadrática $\boldsymbol{\xi}'\mathbb{V}(\boldsymbol{x})\boldsymbol{\xi}$ é não-negativa para todo $\boldsymbol{\xi} \in \mathbb{R}^d$).

2. Mostre que $\mathbb{V}(\boldsymbol{x})$ é singular (isto é, não-invertível) se, e somente se, existem um vetor (não aleatório) $\boldsymbol{\zeta} \in \mathbb{R}^d$ e uma constante $\eta \in \mathbb{R}$ tais que
$$\mathbb{P}(\boldsymbol{\zeta}'\boldsymbol{x} = \eta) = 1.$$

Exercício 11. Seja $\boldsymbol{X}$ uma matriz aleatória de dimensões $n \times (d+1)$. Encontre condições que garantam a invertibilidade de $\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})$.

Esperança condicional e modelos de regressão: um interlúdio

Dante Alighieri, Apoteosi a Firenze, Giovanni Guida

"Ó, vós que entrais, abandonai toda a esperança!"
Dante Alighieri, *Divina Commedia*

Suponha que $\boldsymbol{y}$, $\boldsymbol{X}$ e $\boldsymbol{u}$ satisfazem[23] a equação

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{8}$$

para algum vetor $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$. Se nada mais for dito, não há muito por onde seguir, então vamos acrescentar à supracitada suposição as hipóteses de que $\mathbb{E}(\boldsymbol{u}) = \mathbf{0}$ e $\mathbb{E}(\boldsymbol{X'u}) = \mathbf{0}$. Com isso, podemos apelar ao Teorema 5 e concluir que, necessariamente, vale que

$$\boldsymbol{\beta} = \big(\mathbb{E}(\boldsymbol{X'X})\big)^{-1}\mathbb{E}(\boldsymbol{X'y}). \tag{9}$$

Um erro de interpretação bastante comum é tomar, nas condições acima, como verdadeira a sentença que afirma: "$\boldsymbol{X\beta}$ é a esperança condicional de $\boldsymbol{y}$ dado $\boldsymbol{X}$". Ocorre que essa sentença é, em geral, falsa: essa é uma das conclusões do Teorema 10 logo abaixo. Os exercícios 14 e 22 ilustram casos em que $\mathbb{E}(\boldsymbol{X'u}) = \mathbf{0}$ mas $\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X}) \neq \boldsymbol{X\beta}$.

[23] Claro, a essa altura já sabemos (ou deveríamos saber) de cor e salteado o que os símbolos $\boldsymbol{y}$, $\boldsymbol{X}$ e $\boldsymbol{u}$ representam.

Na literatura de Econometria, convencionou-se chamar de **exogeneidade** à propriedade $\mathbb{E}(\boldsymbol{X'u}) = \mathbf{0}$, e de **exogeneidade estrita** à propriedade $\mathbb{E}(\boldsymbol{u} \,|\, \boldsymbol{X}) \overset{q.c.}{=} \mathbf{0}$. No primeiro caso dizemos também que $\boldsymbol{X}$ é um regressor **exógeno** e, no segundo, que é **estritamente exógeno**.[24] Afortunadamente, o adjetivo "estrito"

[24] As coisas se tornam mais simples quando estamos em um cenário iid, pois aí temos $\mathbb{E}(\boldsymbol{X'u}) = n\mathbb{E}(\boldsymbol{x}_1 u_1)$. Nesse caso, usamos o jargão recém introduzido para referirmo-nos aos vetores $\boldsymbol{x}_i$'s.

aqui não é mero jogo de palavras, já que temos o seguinte:

Teorema 8. Se $\boldsymbol{X}$ é estritamente exógeno, então $\boldsymbol{X}$ é exógeno.

Em vista desse resultado, quando nos deparamos com O Modelo linear (8) e desejamos interpretá-lo como uma relação que descreve a *esperança condicional* de $\boldsymbol{y}$ dado $\boldsymbol{X}$ como uma transformação linear de $\boldsymbol{X}$, necessitamos forçosamente *supor* que $\boldsymbol{X}$ é estritamente exógeno.[25]

[25] Isso é particularmente importante em aplicações, onde é costumeiro dar uma interpretação causal ao modelo linear, tomando o coeficiente β_j como o *efeito* parcial do regressor x_{1j} na resposta y_1.

Projeções: uma primeira incursão

Um **espaço de Hilbert** é uma estrutura matemática (um conjunto) dotada de propriedades algébricas, topológicas e geométricas semelhantes àquelas que são próprias dos espaços Euclidianos. De fato, para todo $d \in \mathbb{N}$ tem-se que $\mathbb{R}^d$ (munido do produto escalar usual) é um espaço de Hilbert — o interessante é que há espaços de Hilbert de dimensão infinita!

Antes de definir o que é um espaço de Hilbert, precisamos da noção de um **produto interno** em um espaço vetorial[26] H, que nada mais é do que uma função $\rho\colon H \times H \to \mathbb{R}$ satisfazendo as seguintes condições, para quaisquer $h, f, g \in H$ e $\xi \in \mathbb{R}$:

PI1. $\rho(h, f) = \rho(f, h)$
PI2. $\rho(\xi h, f) = \xi\rho(h, f)$
PI3. $\rho(h + f, g) = \rho(h, g) + \rho(f, g)$
PI4. $\rho(h, h) \geq 0$
PI5. Se $\rho(h, h) = 0$ então $h = 0 \in H$.

[26] Tomaremos, por simplicidade, espaços vetoriais sobre o corpo dos números reais.

A partir de um produto interno ρ em um espaço vetorial H, é possível introduzir uma noção de *comprimento* de um elemento $h \in H$: escrevemos

$$\|h\|_\rho = \sqrt{\rho(h, h)}$$

e chamamos essa quantidade a **norma** de h (induzida por ρ). Com essa notação, estamos agora aptos a definir espaços de Hilbert, e a definição é a seguinte: dados um espaço vetorial H e um produto interno ρ nesse espaço, chamamos ao par (H, ρ) um **espaço de Hilbert** se valer a seguinte condição:[27]

H1. Se $h_0, h_1, h_2, \ldots$ é uma sequência de elementos de H tal que

$$\sum\nolimits_{k=1}^{\infty} \|h_k - h_{k-1}\|_\rho < \infty,$$

[27] A propriedade *H1* exprime o atributo de *completude* do espaço H: ela significa que, nesse espaço, sequências cujos termos sucessivos tornam-se arbitrariamente próximos uns dos outros necessariamente aproximam-se de algum elemento de H (em certo sentido, isso significa que "não faltam pontos em H" — pense no conjunto $\mathbb{Q}$ dos números racionais e em como podemos aproximar o número irracional $\sqrt{2}$ por elementos de $\mathbb{Q}$).

então existe um elemento $h \in H$ tal que $\lim_{n\to\infty} \|h - h_n\|_\rho = 0$.

Exemplo 9. Considere o conjunto H assim definido: um objeto $\boldsymbol{x}$ pertence a H se, e somente se, $\boldsymbol{x}$ é um vetor aleatório de dimensão d tal que $\mathbb{E}(\boldsymbol{x}'\boldsymbol{x}) < \infty$. Está claro que H é um espaço vetorial. Vamos então definir uma função $\rho\colon H \times H \to \mathbb{R}$, pela expressão

$$\rho(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}(\boldsymbol{x}'\boldsymbol{z}), \qquad \boldsymbol{x}, \boldsymbol{z} \in H. \tag{10}$$

Não é difícil de verificar que ρ assim definida satisfaz as condições *PI1* a *PI4* acima. Quanto a *PI5*, note que se $\mathbb{P}[\boldsymbol{x} = \boldsymbol{0}] = 1$, então $\|\boldsymbol{x}\|_\rho = 0$ e, reciprocamente, se $\|\boldsymbol{x}\|_\rho = 0$ então necessariamente vale que $\mathbb{P}[\boldsymbol{x} = \boldsymbol{0}] = 1$. Por essa razão, identificamos[28] elementos de H que sejam iguais com probabilidade 1. Um fato importante (e, infelizmente, aqui não vamos prová-lo) é que o espaço H é um espaço de Hilbert e, exceto no caso em que Ω é um conjunto finito, tipicamente tem dimensão infinita. Aproveitando a deixa, vamos introduzir uma notação e escrever $H =: L^2_{\mathbb{P}}\big(\Omega; \mathbb{R}^d\big)$.

[28] A matemática aqui tem lá suas nuances: o que estamos fazendo — ávidos por obter a propriedade *PI5* — é o que se chama de *tomar o quociente* do espaço H por uma relação de equivalência, no caso a relação em que dois elementos $\boldsymbol{x}$ e $\boldsymbol{z}$ de H são ditos equivalentes se

$$\mathbb{P}[\boldsymbol{x} = \boldsymbol{z}] = 1.$$

Produtos internos também são úteis, entre outros motivos, por permitirem que em H sejam introduzidas noções de ângulos entre elementos — de particular interesse para nós é a noção de ortogonalidade: dizemos que $h, f \in H$ são **ortogonais** se $\rho(h, f) = 0$. Espaços de Hilbert possuem a importante propriedade de serem *proximinais*, o que quer dizer o seguinte: se $h \in H$ está dado, e se $M \subseteq H$ é um subespaço fechado[29] de H, então existe um único elemento $h_M \in M$ tal que

$$\|h - h_M\|_\rho \leq \|h - f\|_\rho, \tag{11}$$

para qualquer $f \in M$. Em particular, a desigualdade acima vale estritamente se tomarmos $f \neq h_M$. O elemento h_M é dito a **projeção ortogonal** de h em M, e vale a decomposição

[29] M é dito fechado quando vale a seguinte propriedade: se $h_1, h_2, \ldots$ é uma sequência convergente de elementos de M, então o limite é também um elemento de M.

$$\rho(h_M, h - h_M) = 0, \tag{12}$$

ou seja, h_M e $h - h_M$ são ortogonais entre si. Escreveremos $M^\perp$ para denotar o **complemento ortogonal de** M, definido como o conjunto de todos os elementos de H que são ortogonais a M — isto é, $f \in M^\perp$ se, e somente se, $\rho(f, g) = 0$ para todo $g \in M$. Note que a desigualdade (11) nos diz que h_M é o elemento de M que está o mais próximo possível de h.

Um caso de particular interesse para nós ocorre quando o subsepaço M discutido acima tem dimensão finita (digamos, $\dim(M) = d+1$); se esse é o caso, então existem elementos $f_0, f_1, \ldots, f_d \in M$ tais que o conjunto por eles formado é linearmente independente e tal que todo elemento $f \in M$ se escreve unicamente na forma

$$f = a_0 f_0 + a_1 f_1 + \cdots + a_d f_d, \tag{13}$$

para alguma $(d+1)$–upla de números reais $a_0, \ldots, a_d$. Reciprocamente, se $f_0, f_1, \ldots, f_d$ são elementos linearmente independentes em H, então o conjunto formado por todos os elementos $f \in H$ que podem ser expressos na forma (13) tem dimensão finita e igual a $d+1$.[30]

Em suma, um subespaço $M \subseteq H$ tem dimensão finita (e igual a $d+1$) se, e somente se, $M = \text{span}(f_0, \ldots, f_d)$, para algum $d \in \mathbb{N}$ e alguma $(d+1)$–upla $f_0, \ldots, f_d$ de elementos linearmente independentes de H. Se esse é o caso, então $h_M = \sum_{j=0}^{d} a_j f_j$ para alguma única $(d+1)$–upla $a_0, \ldots, a_d \in \mathbb{R}$, e a equação (11) nos diz que

$$\left\| h - \sum_{j=0}^{d} a_j f_j \right\|_\rho \leq \left\| h - \sum_{j=0}^{d} b_j f_j \right\|_\rho,$$

para qualquer $\boldsymbol{b} = (b_0 \quad b_1 \quad \ldots \quad b_d)' \in \mathbb{R}^{d+1}$. Concluímos, então, que o vetor $\boldsymbol{a} = (a_0 \quad a_1 \quad \cdots \quad a_d)' \in \mathbb{R}^{d+1}$ é a única solução para o problema de minimização

$$\min_{\boldsymbol{b} \in \mathbb{R}^{d+1}} \left\| h - \sum_{j=0}^{d} b_j f_j \right\|_\rho$$

Agora que sabemos algo sobre projeções ortogonais, podemos enunciar o seguinte importantíssimo resultado.

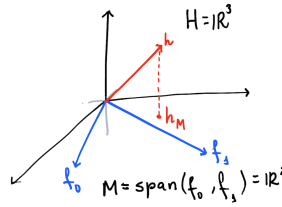


Figura 3: Projeção ortogonal de h no subespaço gerado por f_0 e f_1.

[30] Denotamos tal conjunto por $\text{span}(f_0, \ldots, f_d)$ e dizemos que esse é o **espaço gerado** por $f_0, \ldots, f_d$.

Teorema 10. Sejam

$$\boldsymbol{y}' = (y_1 \quad \dots \quad y_n) \quad e \quad \boldsymbol{x}_i' = (x_{i,0} \quad x_{i,1} \quad \cdots \quad x_{id}),$$

onde $i \in \{1, \dots, n\}$, vetores aleatórios de dimensões n e $d+1$, respectivamente, onde $n \geq d+1$. Seja ainda

$$\boldsymbol{X} = \begin{pmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nd} \end{pmatrix}. \tag{14}$$

Vale o seguinte:

1. No espaço $L^2_{\mathbb{P}}\big(\Omega; \mathbb{R}^1\big)$ com o produto interno

$$\rho(z, w) = \mathbb{E}(zw), \qquad z, w \in L^2_{\mathbb{P}}\big(\Omega; \mathbb{R}^1\big)$$

   (a) a esperança condicional $\mathbb{E}(y_1 \,|\, \boldsymbol{x}_1)$ é a projeção ortogonal da variável aleatória y_1 no subespaço M constituído por todas as variáveis aleatórias z que podem ser representadas na forma $z = g \circ \boldsymbol{x}_1$ para alguma função $g\colon \mathbb{R}^{d+1} \to \mathbb{R}^1$.

   (b) se $\mathbb{E}(\boldsymbol{x}_1 \boldsymbol{x}_1')$ é invertível, então o elemento $\boldsymbol{x}_1' \boldsymbol{\beta}$, com $\boldsymbol{\beta} \coloneqq \big(\mathbb{E}(\boldsymbol{x}_1 \boldsymbol{x}_1')\big)^{-1} \mathbb{E}(\boldsymbol{x}_1 y_1)$, é a projeção ortogonal de y_1 no subespaço $\operatorname{span}(x_{1,0}, \dots, x_{1d})$.

2. No espaço $L^2_{\mathbb{P}}\big(\Omega; \mathbb{R}^n\big)$ com o produto interno

$$\rho(\boldsymbol{z}, \boldsymbol{w}) = \mathbb{E}(\boldsymbol{z}'\boldsymbol{w}), \qquad \boldsymbol{z}, \boldsymbol{w} \in L^2_{\mathbb{P}}\big(\Omega; \mathbb{R}^n\big),$$

   (a) a esperança condicional $\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X})$ é a projeção ortogonal do vetor aleatório $\boldsymbol{y}$ no subespaço M constituído por todos os vetores aleatórios $\boldsymbol{z}$ que podem ser representados na forma $\boldsymbol{z} = g \circ \boldsymbol{X}$ para alguma função $g\colon \mathbb{R}^{n \times (d+1)} \to \mathbb{R}^n$.

   (b) se $\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})$ é invertível, então o elemento $\boldsymbol{X}\boldsymbol{\beta}$, com $\boldsymbol{\beta} \coloneqq \big(\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})\big)^{-1} \mathbb{E}(\boldsymbol{X}'\boldsymbol{y})$, é a projeção ortogonal de $\boldsymbol{y}$ no subespaço gerado pelas $d+1$ colunas de $\boldsymbol{X}$.

3. No espaço $\mathbb{R}^n$ com o produto escalar usual, se $\boldsymbol{X}'\boldsymbol{X}$ é invertível então o elemento $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$, onde $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, é a projeção ortogonal do vetor $\boldsymbol{y}$ no subespaço gerado pelas $d+1$ colunas de $\boldsymbol{X}$.

Demonstração. Vamos demonstrar o item (2a). O item (1a) segue diretamente tomando-se $n = 1$. As demonstrações dos itens (1b), (2b) e (3) são exercícios.

Seja $\varphi\colon \mathbb{R}^{n \times (d+1)} \to \mathbb{R}^n$ a função de regressão de $\boldsymbol{y}$ em $\boldsymbol{X}$, quer dizer $\varphi \circ \boldsymbol{X} = \mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X})$, e seja $g\colon \mathbb{R}^{n \times (d+1)} \to \mathbb{R}^n$ uma função mensurável qualquer satisfazendo $\mathbb{E}(|g \circ \boldsymbol{X}|) < \infty$, onde $|\cdot|$

denota a norma Euclidiana em $\mathbb{R}^n$, $|\boldsymbol{\xi}| \coloneqq \sqrt{\boldsymbol{\xi}'\boldsymbol{\xi}}$. Escrevendo $\boldsymbol{z} = \varphi \circ \boldsymbol{X}$ e $\boldsymbol{w} = g \circ \boldsymbol{X}$, temos

$$\begin{aligned}
\|\boldsymbol{y} - g \circ \boldsymbol{X}\|_\rho^2 &= \mathbb{E}\big\{|\boldsymbol{y} - g \circ \boldsymbol{X}|^2\big\} \\
&= \sum_{i=1}^n \mathbb{E}\Big\{(y_i - w_i)^2\Big\} \\
&= \sum_{i=1}^n \mathbb{E}\Big\{((y_i - z_i) + (z_i - w_i))^2\Big\} \\
&= \sum_{i=1}^n \mathbb{E}\big\{(y_i - z_i)^2 + 2(y_i - z_i)(z_i - w_i) + (z_i - w_i)^2\big\}.
\end{aligned}$$

Distribuindo, por linearidade, o operador de esperança e notando que

$$\begin{aligned}
\mathbb{E}\{(y_i - z_i)(z_i - w_i)\} &= \mathbb{E}\{\mathbb{E}[(y_i - z_i)(z_i - w_i) \,|\, \boldsymbol{X}]\} \\
&= \mathbb{E}\{(z_i - w_i)\mathbb{E}[(y_i - z_i) \,|\, \boldsymbol{X}]\} \\
&= 0,
\end{aligned}$$

ficamos com

$$\begin{aligned}
\|\boldsymbol{y} - g \circ \boldsymbol{X}\|_\rho^2 &= \sum_{i=1}^n \mathbb{E}\big\{(y_i - z_i)^2\big\} + \sum_{i=1}^n \mathbb{E}\big\{(z_i - w_i)^2\big\} \\
&= \|\boldsymbol{y} - \varphi \circ \boldsymbol{X}\|_\rho^2 + \|\varphi \circ \boldsymbol{X} - g \circ \boldsymbol{X}\|_\rho^2.
\end{aligned}$$

O primeiro termo no lado direito da igualdade acima independe de g; o segundo é não negativo e se anula precisamente e tão somente quando tomamos g tal que $\mathbb{P}[g \circ \boldsymbol{X} = \varphi \circ \boldsymbol{X}] = 1$. Isso completa a demonstração. ■

Remark 11. Outra maneira de enunciar, digamos, o item (1b) do teorema acima é a seguinte: de todas as variáveis aleatórias z que podem ser escritas na forma

$$z = b_0 x_{1,0} + \cdots + b_d x_{1,d} = \boldsymbol{x}_1'\boldsymbol{b},$$

para algum $\boldsymbol{b} \in \mathbb{R}^{d+1}$, temos que $\boldsymbol{x}_1'\boldsymbol{\beta}$, com $\boldsymbol{\beta} \coloneqq \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1)$, é aquela que está mais próxima de y_1 no sentido de que

$$\mathbb{E}\big\{(y_1 - \boldsymbol{x}_1'\boldsymbol{\beta})^2\big\} \leq \mathbb{E}\big\{(y_1 - \boldsymbol{x}_1'\boldsymbol{b})^2\big\}$$

para todo $\boldsymbol{b} \in \mathbb{R}^{d+1}$, em que a desigualdade vale estritamente se tomarmos $\boldsymbol{b} \neq \boldsymbol{\beta}$. Para fins mnemônicos, convém reescrever cada um dos demais itens do teorema de forma semalhante ao que fizemos aqui.

Exercícios

Quaisquer erros nos enunciados são — é claro! — intencionais.

Exercício 12. Nesse capítulo utilizamos a notação $\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X})$ sem fornecer uma definição. Ei-la: sejam $\boldsymbol{y}$ um vetor aleatório de dimensão n, sejam $\boldsymbol{x}_i$ vetores aleatórios de dimensão $d+1$, onde $i \in \{1, \ldots, n\}$, e seja $\boldsymbol{X}$ definida como em (14). A **esperança condicional** de $\boldsymbol{y}$ dado $\boldsymbol{X}$ (equivalentemente, dados $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$) é o vetor aleatório $\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X}) \equiv \mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ de dimensão n que satisfaz as seguintes condições:

EC1. $\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X}) = \varphi \circ \boldsymbol{X}$, para alguma função mensurável $\varphi\colon \mathbb{R}^{n\times(d+1)} \to \mathbb{R}^n$, com $\mathbb{E}(|\varphi \circ \boldsymbol{X}|) < \infty$.

EC2. Vale a igualdade $\mathbb{E}\big(\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X})\, \mathbb{I}_{[\boldsymbol{X} \in B]}\big) = \mathbb{E}(\boldsymbol{y}\, \mathbb{I}_{[\boldsymbol{X} \in B]})$ para todo subconjunto Boreliano $B \subseteq \mathbb{R}^{n\times(d+1)}$.

1. Observando que qualquer função $g\colon \mathbb{R}^m \to \mathbb{R}^k$ é da forma $g = (g_1, \ldots, g_k)$, onde $g_i\colon \mathbb{R}^m \to \mathbb{R}$, verifique que
$$\big[\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X})\big]_i = \mathbb{E}(y_i \,|\, \boldsymbol{X}), \quad i \in \{1, \ldots, n\}$$
2. Mostre que, se z é uma variável aleatória da forma $z = g \circ \boldsymbol{X}$ para alguma função g de $\mathbb{R}^{n\times(d+1)}$ em $\mathbb{R}$, então $\mathbb{E}(z\boldsymbol{y} \,|\, \boldsymbol{X}) = z\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X})$.

Exercício 13. Mostre que, se os vetores aleatórios
$$(\boldsymbol{x}_1 \quad y_1)', \ldots, (\boldsymbol{x}_n \quad y_n)'$$
formam uma amostra aleatória, então $\big[\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X})\big]_i = \mathbb{E}(y_i \,|\, \boldsymbol{x}_i)$.

Exercício 14. Suponha que $x_{1,1}, \ldots, x_{n,1}$ e $v_1, \ldots, v_n$ são uma amostra aleatória de x_1 e v, respectivamente, onde x_1 e v são como no exercício 4: mutuamente independentes e, cada um, simétrico em torno de zero. Para $i \in \{1, \ldots, n\}$, defina $y_i = x_{i1}^2 + v_i$, e ponha
$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} \\ \vdots & \vdots \\ 1 & x_{n,1} \end{pmatrix}.$$
Sendo $\boldsymbol{\beta}$ definido como na equação (9), e pondo $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$, já sabemos que $\mathbb{E}(\boldsymbol{X}'\boldsymbol{u}) = 0$ e, como as variáveis aleatórias aqui são identicamente distribuídas, sabemos de fato que $\mathbb{E}(\boldsymbol{u}) = 0$.

1. Mostre que $\mathbb{E}(\boldsymbol{u} \,|\, \boldsymbol{X}) \neq 0$ e conclua que $\mathbb{E}(\boldsymbol{y} \,|\, \boldsymbol{X}) \neq \boldsymbol{X}\boldsymbol{\beta}$.

2. Defina agora, para $i \in \{1, \ldots, n\}$, $x_{i2} = x_{i1}^2$ e ponha

$$\boldsymbol{W} \coloneqq \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix}, \quad \boldsymbol{\gamma} \coloneqq \big(\mathbb{E}(\boldsymbol{W'W})\big)^{-1}\mathbb{E}(\boldsymbol{W'y}).$$

   Mostre que $\boldsymbol{\gamma}' = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$ e $\mathbb{E}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\gamma} \,|\, \boldsymbol{W}) = \boldsymbol{0}$.

Exercício 15. Demonstre o Teorema 8.

Exercício 16. Seja $M(n \times d)$ o conjunto formado por todas as matrizes $n \times d$, e considere a transformação $\tau\colon M(n \times d) \to \mathbb{R}^{n \times d}$ que "empilha" as colunas de uma matriz $A \in M(n \times d)$ para obter um vetor em $\mathbb{R}^{n \times d}$. Verifique que τ é linear, bijetora, contínua e que τ^{-1} é também contínua.[31] Esse exercício justifica a notação utilizada para o domínio da função de regressão de $\boldsymbol{y}$ em $\boldsymbol{X}$.

[31] Sobre continuidade: é preciso verificar que, se $A_k \to A$ em $M(n \times d)$, então $\tau A_k \to \tau A$ em $\mathbb{R}^{n \times d}$ etc. Em ambos os casos, convergência significa convergência componente-a-componente.

Exercício 17. Considere uma matriz $A \in M(2 \times 2)$,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

e seja $\boldsymbol{a} = \tau A \in \mathbb{R}^4$, onde τ é definida como no exercício acima. Sejam ainda $\boldsymbol{\beta} \in \mathbb{R}^2$ e $B \in M(2 \times 4)$. Comparando os vetores $A\boldsymbol{\beta}$ e $B\boldsymbol{a}$, conclua que nem toda transformação linear $T\colon M(2 \times 2) \to \mathbb{R}^2$ é da forma $A \mapsto A\boldsymbol{\beta}$ para algum $\boldsymbol{\beta} \in \mathbb{R}^2$.

Exercício 18. Verifique que o plano Euclidiano $\mathbb{R}^2$, munido do produto escalar usual, é um espaço de Hilbert.

Exercício 19. Mostre que o espaço $L^2_{\mathbb{P}}\big(\Omega; \mathbb{R}^1\big)$ coincide com o conjunto formado por todas as variáveis aleatórias com variâncias finitas.

Exercício 20. Dado um subespaço fechado M em um espaço de Hilbert H, mostre que todo elemento $h \in H$ se exprime de maneira única como $h = h_M + h_{M^\perp}$.

Exercício 21. Demonstre os itens (1b), (2b) e (3) do Teorema 10. Dica: comece pelo item (3), notando que a função real L definida,

para $\boldsymbol{b} \in \mathbb{R}^{d+1}$, por

$$L(\boldsymbol{b}) \coloneqq |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}|^2 = \sum_{i=1}^{n} \big(y_i - [\boldsymbol{X}\boldsymbol{b}]_i\big)^2,$$

é uma função convexa. Isso permite que encontremos o valor $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{d+1}$ que minimiza globalmente a função L resolvendo o sistema

$$\nabla L(\widehat{\boldsymbol{\beta}}) = \mathbf{0},$$

onde ∇L denota o gradiente de L, $[\nabla L]_j \coloneqq \partial L/\partial b_j$, $j \in \{0, \dots, d\}$.

Para o item (2b), basta seguir um roteiro semelhante ao acima, agora definindo $L(\boldsymbol{b}) = \mathbb{E}\big\{|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}|^2\big\}$ para $\boldsymbol{b} \in \mathbb{R}^{d+1}$, onde $|\cdot|$ denota a norma Euclidiana, e resolver a equação $\nabla L(\boldsymbol{\beta}) = \mathbf{0}$. Sinta-se autorizado a fazer a passagem $\nabla\mathbb{E} = \mathbb{E}\nabla$.

Exercício 22. Seja $(\Omega, \mathscr{A}, \mathbb{P})$ o espaço de probabilidade em que $\Omega = (0, 1]$, $\mathscr{A}$ = 'σ-álgebra de Borel em $(0, 1]$' e $\mathbb{P}$ = 'medida de Lebesgue em $(0, 1]$' (isto é, $\mathbb{P}(a, b] = b - a$ para qualquer intervalo $(a, b] \subseteq (0, 1]$, $0 \leq a < b \leq 1$). Seja x a variável aleatória definida via $x(\omega) \coloneqq \omega$, $\omega \in \Omega$, e seja y uma variável aleatória integrável qualquer.

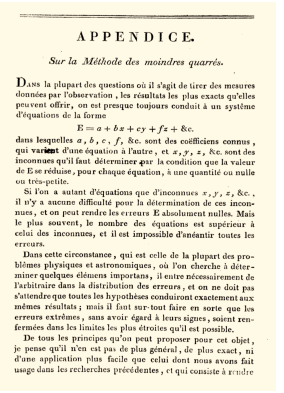
1. Mostre que $\mathbb{E}(y \,|\, x) = y$.
2. Sendo $M = \operatorname{span}(x) \subseteq L^2_{\mathbb{P}}(\Omega; \mathbb{R})$, mostre que $y_M = 3x\,\mathbb{E}(xy)$.

O método de mínimos quadrados

> "We all believe fairy-tales, and live in them. Some, with a sumptuous literary turn, believe in the existence of the lady clothed with the sun. Some, with a more rustic, elvish instinct, believe merely in the impossible sun itself."
>
> G. K. Chesterton, Heretics

A presente seção é um ínterim, durante o qual poderemos esquecer — provisoriamente — que os objetos com os quais estamos lidando são variáveis (e vetores e matrizes) aleatórias. Sendo assim, por ora $\boldsymbol{y}$ denota um vetor de $\mathbb{R}^n$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ vetores de $\mathbb{R}^{d+1}$ e $\boldsymbol{X}$ uma matriz $n \times (d+1)$ cuja componente i, j é $x_{ij} = j$ésima componente de $\boldsymbol{x}_i$. Tipicamente (embora não necessariamente) convenciona-se que $x_{i0} = 1$ para todo i.

O MÉTODO DOS MÍNIMOS QUADRADOS APARECE PELA PRIMEIRA VEZ EM 1805, em um apêndice intitulado *Sur la Méthode des moindres quarrés*, na *Nouvelles méthodes pour la détermination des orbites des comètes* de Adrien–Marie Legendre. Em 1809, o mesmo método aparece na *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium* de Carl Friedrich Gauss. Tudo indica que Gauss e Legendre "descobriram" o método de forma independente, e há uma contenda a respeito de quem, de fato, foi o primeiro descobridor (foi Gauss).[32] O referido método pode ser subdivido em duas categorias, a saber, mínimos quadrados *ordinários* e mínimos quadrados *não-lineares*, e por ora estaremos interessados na primeira delas. Aqui, o termo "ordinários" é empregado em contraposição ao termo "não-lineares" e, portanto, um nome mais adequado (ou ao menos mais explícito) para a primeira categoria seria *mínimos quadrados lineares*. Como veremos, *linea-*



APPENDICE.

Sur la Méthode des moindres quarrés.

DANS la plupart des questions où il s'agit de tirer des mesures données par l'observation, les résultats les plus exacts qu'elles peuvent offrir, on est presque toujours conduit à un système d'équations de la forme

$$E = a + bx + cy + fz + \&c.$$

dans lesquelles $a, b, c, f,$ &c. sont des coëfficiens connus, qui varient d'une équation à l'autre, et $x, y, z,$ &c. sont des inconnues qu'il faut déterminer par la condition que la valeur de E se réduise, pour chaque équation, à une quantité ou nulle ou très-petite.

Si l'on a autant d'équations que d'inconnues $x, y, z,$ &c., il n'y a aucune difficulté pour la détermination de ces inconnues, et on peut rendre les erreurs E absolument nulles. Mais le plus souvent, le nombre des équations est supérieur à celui des inconnues, et il est impossible d'anéantir toutes les erreurs.

Dans cette circonstance, qui est celle de la plupart des problèmes physiques et astronomiques, où l'on cherche à déterminer quelques élémens importans, il entre nécessairement de l'arbitraire dans la distribution des erreurs, et on ne doit pas s'attendre que toutes les hypothèses conduiront exactement aux mêmes résultats; mais il faut sur-tout faire en sorte que les erreurs extrêmes, sans avoir égard à leurs signes, soient renfermées dans les limites les plus étroites qu'il est possible.

De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre

A primeira página do apêndice *Sur la Méthode des moindres quarrés*. Legendre vai direto ao ponto: "Dans la plupart des questions où il s'agit de tirer des mesures données par l'observation, les résultats les plus exacts qu'elles peuvent offrir, on est presque toujours conduit à un systême d'équations..."

[32] R.L. Plackett. The discovery of the method of least squares. *Biometrika*, 59(2), 1972

ridade e *não-linearidade* referem-se à maneira como os *parâmetros* se apresentam nos sistemas de equações estudados.[33]

[33] Veja os exercícios 32 e 33.

O método

Sem mais delongas, consideremos como dados n pontos[34] no espaço Euclidiano $\mathbb{R}^{d+2}$:

$$\begin{pmatrix} y_1 \\ x_{1,0} \\ \vdots \\ x_{1,d} \end{pmatrix}, \begin{pmatrix} y_2 \\ x_{2,0} \\ \vdots \\ x_{2,d} \end{pmatrix}, \dots, \begin{pmatrix} y_n \\ x_{n,0} \\ \vdots \\ x_{n,d} \end{pmatrix}.$$

[34] Em que pese a crescente importância de se considerar situações onde $n < d+1$ — por exemplo, na literatura de regularização e seleção de variáveis — por ora vamos nos restringir ao tradicionalíssimo cenário em que $n > d+1$. Tipicamente, cada um desses n pontos representa uma *observação* de $d+2$ variáveis.

Para cada vetor $\boldsymbol{b} = (b_0 \quad b_1 \quad \cdots \quad b_d)' \in \mathbb{R}^{d+1}$, podemos escrever um sistema de n equações lineares,

$$\begin{aligned} y_1 &= b_0 x_{1,0} + b_1 x_{1,1} + \cdots + b_d x_{1d} + v_1(\boldsymbol{b}) \\ y_2 &= b_0 x_{2,0} + b_1 x_{2,1} + \cdots + b_d x_{2d} + v_2(\boldsymbol{b}) \\ &\vdots \\ y_n &= b_0 x_{n0} + b_1 x_{n1} + \cdots + b_d x_{nd} + v_n(\boldsymbol{b}) \end{aligned} \tag{15}$$

onde os termos $v_i(\boldsymbol{b})$, $i \in \{1, \dots, n\}$, estão definidos, por tautologia, como

$$v_i(\boldsymbol{b}) \coloneqq y_i - \boldsymbol{x}_i'\boldsymbol{b}, \quad i \in \{1, \dots, n\}.$$

Visto que estamos considerando $n > d+1$, ocorre que — exceto em casos mui particulares — não é possível encontrar um vetor $\boldsymbol{b}^* \in \mathbb{R}^{d+1}$ tal que valha $v_i(\boldsymbol{b}^*) = 0$ para todo i. Como prosseguir, então? Ora, eis a questão: tentemos encontrar um vetor $\widehat{\beta} \in \mathbb{R}^{d+1}$ que torne todos os **resíduos** $\widehat{u}_i \coloneqq v_i(\widehat{\boldsymbol{\beta}})$, $i \in \{1, \dots n\}$, "tão pequenos quanto for possível". Claro, há infinitas maneiras de se dar sentido à expressão "tão pequenos quanto for possível" — Boscovich e Laplace que o digam! — e a grande sacada de Legendre e Gauss foi considerar que o vetor $\widehat{\boldsymbol{\beta}}$ desejado deve ser aquele que soluciona o problema de se minimizar, com respeito a $\boldsymbol{b} \in \mathbb{R}^{d+1}$, a expressão

$$\sum_{i=1}^{n} v_i(\boldsymbol{b})^2 \equiv \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{b})^2 \equiv (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \equiv \left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\right|^2,$$

onde $|\cdot|$ aqui denota a norma Euclidiana em $\mathbb{R}^n$. Isto é,

$$\widehat{\boldsymbol{\beta}} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^{d+1}} \left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\right|^2. \tag{16}$$

Felizmente a essa altura já estamos familiarizados com espaços de Hilbert, e sabemos que $\mathbb{R}^n$ é um deles! *Ergo*, pela propriedade de proximinalidade, existe um único elemento $\widehat{\boldsymbol{y}}$ pertencente ao espaço gerado pelas colunas de $\boldsymbol{X}$ tal que

$$\left|\boldsymbol{y}-\widehat{\boldsymbol{y}}\right| \leq \left|\boldsymbol{y}-\boldsymbol{\xi}\right|, \quad \boldsymbol{\xi} \in \mathbb{R}^n.$$

Pelo Teorema 10 (e também pelo exercício 21), sabemos que — desde que a matriz $\boldsymbol{X}'\boldsymbol{X}$ seja invertível — o vetor $\widehat{\boldsymbol{y}}$ é dado por

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}, \quad \text{com } \widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Note que isso nos diz quem é a solução na equação (16)! As componentes do vetor $\widehat{\boldsymbol{y}} = (\widehat{y}_1 \quad \cdots \quad \widehat{y}_n)'$ são ditas os **valores ajustados** (por mínimos quadrados ordinários). Explicitamente,

$$\widehat{y}_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} = \sum_{j=0}^{d} \widehat{\beta}_j x_{ij}, \quad i \in \{1,\ldots,n\}.$$

Lembrando que os *resíduos* foram definidos como $\widehat{u}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} = y_i - \widehat{y}_i$, e pondo $\widehat{\boldsymbol{u}} = (\widehat{u}_1 \quad \cdots \quad \widehat{u}_n)'$, podemos reescrever o sistema de equações (15) — agora com $\widehat{\boldsymbol{\beta}}$ no lugar de $\boldsymbol{b}$ — como

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}} = \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{u}}. \tag{17}$$

A igualdade acima nada tem que ver, a princípio, com o modelo linear: não se trata de um modelo, mas de uma relação algébrica sempre válida,[35] relacionando os objetos que ali aparecem. De fato, temos o seguinte:

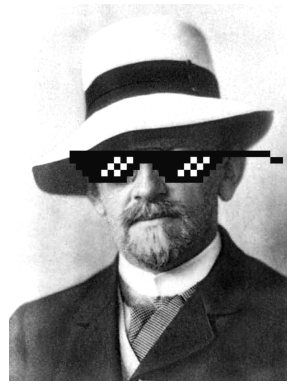
Teorema 12. Seja $M = \mathrm{col}(\boldsymbol{X}) \subseteq \mathbb{R}^n$ o espaço gerado pelas colunas[36] de $\boldsymbol{X}$, e suponha que $\boldsymbol{X}'\boldsymbol{X}$ é invertível. Então $\widehat{\boldsymbol{y}} = \boldsymbol{y}_M$ e $\widehat{\boldsymbol{u}} = \boldsymbol{y}_{M^\perp}$.

Demonstração. Não há o que demonstrar: já vimos[37] que $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ é a projeção ortogonal de $\boldsymbol{y}$ sobre $\mathrm{col}(\boldsymbol{X})$, e já sabemos[38] que a decomposição $\boldsymbol{y} = \boldsymbol{y}_M + \boldsymbol{y}_{M^\perp}$ é única. ∎

Corolário 13. Nas condições do Teorema 12, vale que

$$\boldsymbol{X}'\widehat{\boldsymbol{u}} = \boldsymbol{0} \in \mathbb{R}^{d+1}.$$

Adicionalmente, se $\boldsymbol{X}$ possui uma coluna de 1's, então $\sum_{i=1}^{n} \widehat{u}_i = 0$.



David Hilbert, *circa* 1912.

[35] Desde que $\boldsymbol{X}'\boldsymbol{X}$ seja invertível, de forma a garantir que $\widehat{\boldsymbol{\beta}}$ esteja bem definido.

[36] Observe que aproveitamos a deixa para introduzir aqui a notação $\mathrm{col}(\boldsymbol{X})$.

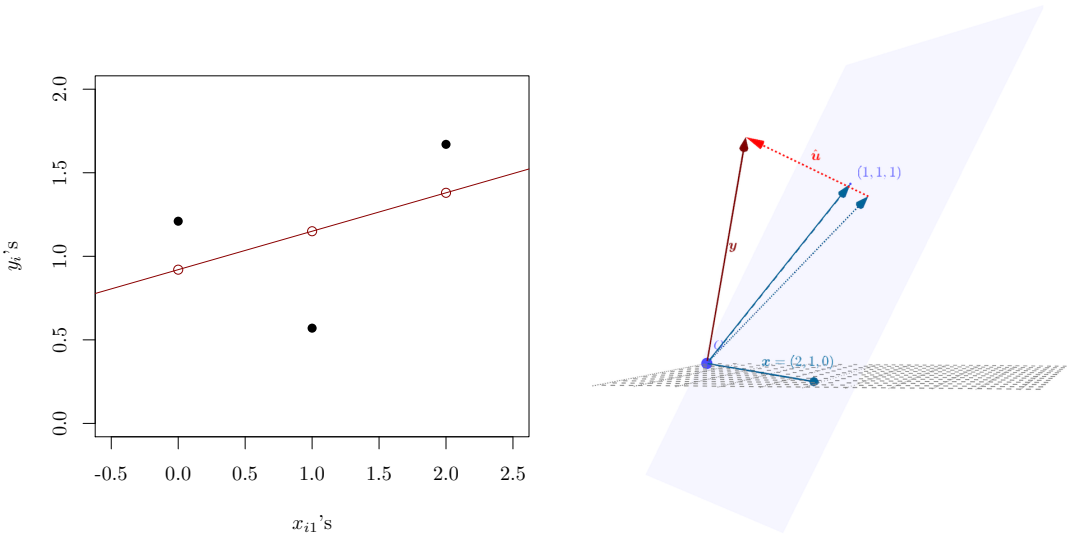
[37] Teorema 10.

[38] Exercício 20.

Exemplo 14. Para ilustrar, consideremos (com $n = 3$ e $d = 1$) o seguinte conjunto de dados:

$$\boldsymbol{y}' = \begin{pmatrix} 1.67 & 0.57 & 1.21 \end{pmatrix}$$
$$\boldsymbol{X}' = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 0 \end{pmatrix}$$

Fazendo as contas encontramos $\widehat{\boldsymbol{\beta}} = (0.92 \quad 0.23)'$ e $\widehat{\boldsymbol{y}} = (1.38 \quad 1.15 \quad 0.92)'$.



As duas figuras acima ilustram duas maneiras distintas de se visualizar os objetos envolvidos no problema. Na figura da esquerda, cada um dos $n = 3$ pontos em preto representa um dos pares ordenados (x_{i1}, y_i). A diagonal em vermelho é a reta dada pela equação $\ell(\xi) = \widehat{\beta}_0 + \widehat{\beta}_1\, \xi$, $\xi \in \mathbb{R}$. Os três círculos situados sobre a reta representam os pares ordenados $(x_{i1}, \widehat{y}_i)$. A figura da direita representa os vetores envolvidos como pontos em $\mathbb{R}^n$, isto é, em $\mathbb{R}^3$. Aqui, cada uma das duas colunas da matriz $\boldsymbol{X}$ está representada por um vetor em azul em linha sólida. O hiperplano (subespaço) gerado pelas colunas de $\boldsymbol{X}$ aparece em azul–claro; o vetor $\boldsymbol{y}$, em vermelho–escuro, "aponta para fora" desse subespaço. O vetor de valores ajustados $\widehat{\boldsymbol{y}}$ está representado pelo vetor em azul em linha pontilhada.

O Teorema de Frisch–Waugh–Lovell

Considere o seguinte *setup*: com $d \geq 1$ — de modo a garantir que a matriz $\boldsymbol{X}$ tenha pelo menos 2 colunas — tomemos inteiros $d_1 > 0$ e

$d_2 > 0$ tais que $d_1 + d_2 = d+1$. Denotemos por $\boldsymbol{X}_1$ a matriz formada pelas d_1 primeiras colunas de $\boldsymbol{X}$ e, semelhantemente, por $\boldsymbol{X}_2$ a matriz das d_2 colunas remanescentes. Simbolicamente, podemos escrever

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix}.$$

Denotemos ainda por $\Pi_1^\perp$ a matriz $(n \times n)$ de projeção ortogonal[39] sobre o complemento ortogonal do espaço gerado pelas colunas de $\boldsymbol{X}_1$. Isto é, assumindo que $\boldsymbol{X}_1'\boldsymbol{X}_1$ é invertível, definimos $\Pi_1^\perp \coloneqq \mathbf{I} - \boldsymbol{X}_1\big(\boldsymbol{X}_1'\boldsymbol{X}_1\big)^{-1}\boldsymbol{X}_1'$, onde $\mathbf{I}$ denota a matriz identidade $n \times n$. *Last but not least*, lembrando que $\widehat{\boldsymbol{\beta}} = \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{y}$, escrevamos

$$\widehat{\boldsymbol{\beta}}' = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_1' & \widehat{\boldsymbol{\beta}}_2' \end{bmatrix}$$

onde $\widehat{\boldsymbol{\beta}}_1$ é o vetor $d_1 \times 1$ cujas componentes correspondem às primeiras d_1 componentes de $\widehat{\boldsymbol{\beta}}$ e assim por diante.

[39] Veja o exercício 29.

Informalmente, o Teorema de Frisch–Waugh–Lovell nos diz que, para computar $\widehat{\boldsymbol{\beta}}_2$, podemos seguir a receita abaixo:

1. "Regrida" $\boldsymbol{y}$ em $\boldsymbol{X}_1$ e obtenha resíduos $\widehat{\boldsymbol{u}}_1 \coloneqq \Pi_1^\perp \boldsymbol{y}$.
2. "Regrida" cada coluna de $\boldsymbol{X}_2$ em $\boldsymbol{X}_1$ e obtenha uma "matriz de resíduos" $\widehat{\boldsymbol{U}}_2 \coloneqq \Pi_1^\perp \boldsymbol{X}_2$.
3. "Regrida" $\widehat{\boldsymbol{u}}_1$ em $\widehat{\boldsymbol{U}}_2$ para obter $\widehat{\boldsymbol{\beta}}_2$.

O enunciado formal é o seguinte:

Teorema 15 (Frisch–Waugh–Lovell)**.** Se $\boldsymbol{X}'\boldsymbol{X}$ é invertível, então

$$\widehat{\boldsymbol{\beta}}_2 = \big(\boldsymbol{X}_2'\Pi_1^\perp\boldsymbol{X}_2\big)^{-1}\boldsymbol{X}_2'\Pi_1^\perp\boldsymbol{y}.$$

Remark 16. O costume nos concede que, ao escrever a solução

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

do problema de minimização (16), seja omitida a dependência — muito evidente — que o termo no lado esquerdo dessa identidade tem nos termos que aparecem à direita. Ora, claramente $\widehat{\boldsymbol{\beta}}$ é uma função de $\boldsymbol{y}$ e $\boldsymbol{X}$, e portanto podemos — quando o clamor por clareza assim demandar — escrever, e.g., $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$. Com essa notação[40] — e notando que $\Pi_1^\perp$ é simétrica e idempotente — o Teorema de Frisch–Waugh–Lovell nos revela que

$$\widehat{\boldsymbol{\beta}}\big(\Pi_1^\perp\boldsymbol{y}, \Pi_1^\perp\boldsymbol{X}_2\big) = \begin{bmatrix} \mathbf{0}_{(d_2\times d_1)} & \mathbf{I}_{(d_2\times d_2)} \end{bmatrix}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}).$$

[40] Por sinal, escrever $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ é uma maneira de dar sentido ao jargão "regredir $\boldsymbol{y}$ em $\boldsymbol{X}$."

Demonstração. Seja $\Pi^\perp = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ a projeção ortogonal sobre $\mathrm{col}(\boldsymbol{X})^\perp$. Notando que $\widehat{\boldsymbol{u}} = \Pi^\perp \boldsymbol{y}$, podemos escrever

$$\boldsymbol{y} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \Pi^\perp \boldsymbol{y}.$$

Isso nos dá, multiplicando a expressão acima, à esquerda, por $\boldsymbol{X}_2'\Pi_1^\perp$,

$$\boldsymbol{X}_2'\Pi_1^\perp \boldsymbol{y} = \boldsymbol{X}_2'\Pi_1^\perp \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2'\Pi_1^\perp \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \boldsymbol{X}_2'\Pi_1^\perp\Pi^\perp \boldsymbol{y}.$$

Agora, o primeiro termo no lado direito da igualdade acima se anula, pois $\Pi_1^\perp \boldsymbol{X}_1 = \mathbf{0}$ (por quê?). De maneira análoga, o terceiro termo se anula: seu transposto é $\boldsymbol{y}'\Pi^\perp\Pi_1^\perp \boldsymbol{X}_2$, e valem as igualdades $\Pi^\perp\Pi_1^\perp = \Pi^\perp$ e $\Pi^\perp \boldsymbol{X}_2 = \mathbf{0}$ (por quê?).[41] Disso segue que $\boldsymbol{X}_2'\Pi_1^\perp \boldsymbol{y} = \boldsymbol{X}_2'\Pi_1^\perp \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2$ e a partir daí obtemos a identidade desejada. ■

[41] Esses "por quês?" não são perguntas retóricas: faça o exercício 29!

Exercícios

Quaisquer erros nos enunciados são — *por supuesto*! — intencionais.

Exercício 23. Mostre cuidadosamente que qualquer solução para o problema

$$\min_{\boldsymbol{b}\in\mathbb{R}^{d+1}} \left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\right|$$

é também uma solução para o problema

$$\min_{\boldsymbol{b}\in\mathbb{R}^{d+1}} \left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\right|^2$$

e vice-versa.

Exercício 24. Encontre condições suficientes e necessárias para que o sistema de equações (15) admita uma solução $\boldsymbol{b}^*$ tal que $v_i(\boldsymbol{b}^*) = 0$ para todo i.

Exercício 25. Encontre condições suficientes e necessárias para que a matriz $\boldsymbol{X}'\boldsymbol{X}$ seja invertível.

Exercício 26. Calcule manualmente o valor de $\widehat{\boldsymbol{\beta}}$ no exemplo 14. Quem são os vetores $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ e $\boldsymbol{x}_3$ nesse exemplo?

Exercício 27. Demonstre o Corolário 13.

Exercício 28 (Representações alternativas)**.** Verifique as seguintes igualdades:

1. $\widehat{\boldsymbol{\beta}} = \big(\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i'\big)^{-1} \sum_{i=1}^n \boldsymbol{x}_i y_i$.
2. $\sum_{i=1}^n \boldsymbol{x}_i u_i = \mathbf{0}$

Exercício 29. Considere as matrizes Π e $\Pi^\perp$ (ambas $n \times n$) definidas por

$$\Pi = \boldsymbol{X}\big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}', \quad \Pi^\perp = \mathbf{I} - \Pi,$$

onde $\mathbf{I}$ denota a matriz identidade (quais as dimensões?). Mostre que:

1. Π e $\Pi^\perp$ são simétricas e idempotentes.[42]
2. Para qualquer vetor $\boldsymbol{\xi} \in \mathrm{col}(\boldsymbol{X})$, vale que $\Pi\boldsymbol{\xi} = \boldsymbol{\xi}$ e $\Pi^\perp\boldsymbol{\xi} = \mathbf{0}$. Em particular, $\Pi\boldsymbol{X} = \boldsymbol{X}$ e $\Pi^\perp\boldsymbol{X} = \mathbf{0}$ (matriz de zeros, $n \times (d+1)$).
3. $\Pi\boldsymbol{y} = \widehat{\boldsymbol{y}}$ e $\Pi^\perp\boldsymbol{y} = \widehat{\boldsymbol{u}}$.
4. (Teorema de Pitágoras) $|\boldsymbol{y}|^2 = |\widehat{\boldsymbol{y}}|^2 + |\widehat{\boldsymbol{u}}|^2$.
5. Supondo agora que $\boldsymbol{X}$ é da forma $\boldsymbol{X} = [\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$, sejam $\Pi_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ e $\Pi_1^\perp = \mathbf{I} - \Pi_1$. Mostre que $\Pi_1 \cdot \Pi = \Pi \cdot \Pi_1 = \Pi_1$ e $\Pi_1^\perp\Pi^\perp = \Pi^\perp\Pi_1^\perp = \Pi^\perp$.
6. Dê os detalhes da demonstração do Teorema de Frisch–Waugh–Lovell.

[42] Simetria significa, e.g., $\Pi = \Pi'$ e idempotência, por seu turno, $\Pi \cdot \Pi = \Pi$.

Exercício 30. Seja $\bar{y}_n \coloneqq n^{-1}(y_1 + \cdots + y_n)$. Verifique as seguintes propriedades:

1. $n^{-1}(\widehat{y}_1 + \cdots + \widehat{y}_n) = \bar{y}_n$.
2. se $d = 0$ e $x_{i0} = 1$, $i \in \{1, \ldots, n\}$, então $\widehat{\boldsymbol{\beta}} = \bar{y}_n$.
3. se $d \geq 1$ e $\boldsymbol{X}'\boldsymbol{X}$ é invertível, então

$$\sum_{i=1}^n (y_i - \bar{y}_n)^2 = \sum_{i=1}^n \{(\widehat{y}_i - \bar{y}_n)^2 + \widehat{u}_i^2\}.$$

Exercício 31. O **coeficiente de determinação** é o número $R^2 \in [0, 1]$ definido por

$$R^2 \coloneqq 1 - \frac{\sum_{i=1}^n \widehat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2}$$

Mostre que

$$R^2 = \frac{\sum_{i=1}^n (\widehat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2}.$$

Exercício 32. Considere uma transformação (possivelmente não-linear) $\Phi\colon \mathbb{R}^{d+1} \to \mathbb{R}^k$. Para $i \in \{1, \dots, n\}$, sejam $\boldsymbol{z}_i \coloneqq \Phi(\boldsymbol{x}_i)$, e $\boldsymbol{Z}$ a matriz cuja entrada i, j é z_{ij}. Interprete o vetor $\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$.

Exercício 33 (Computacional)**.** Com a mesma notação do exercício anterior, seja $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^3$ definida por

$$\Phi(\xi_0, \xi_1) = (\xi_0, \xi_1, \xi_1^3), \quad \boldsymbol{\xi} \in \mathbb{R}^2.$$

1. Utilize o software `RStudio` para gerar uma amostra de tamanho $n = 200$ do modelo

   $$y_i = 1 + x_{i1} + x_{i1}^3 + u_i,$$

   onde x_{i1} tem distribuição uniforme no intervalo $[-2, 2]$ e u_i tem distribuição Normal$(0, \sigma^2)$, com $\sigma = 0.03$, independente de x_{i1}.

2. Faça o gráfico de dispersão (no plano) dos pares (x_{i1}, y_i) simulados e compare-o com o gráfico de dispersão (em $\mathbb{R}^3$) das ternas (x_{i1}, x_{i1}^3, y_i).

3. No gráfico 2D, acrescente a reta cuja equação é

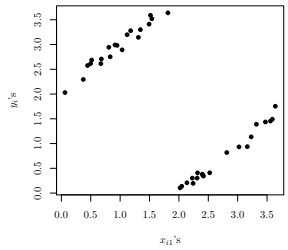
   $$\ell(\xi) = \widehat{\beta}_0 + \widehat{\beta}_1 \xi, \quad \xi \in \mathbb{R}.$$

4. No gráfico 2D, acrescente a cúbica cuja equação é

   $$q(\xi) = \widehat{\gamma}_0 + \widehat{\gamma}_1 \xi_1 + \widehat{\gamma}_2 \xi_1^3, \quad \xi \in \mathbb{R},$$

   onde $\widehat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$, com $\boldsymbol{Z}$ definido como no exercício 32.

Exercício 34. Nesse exercício, vamos escrever $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ para explicitar a dependência da solução (16) no vetor $\boldsymbol{y}$ e na matriz $\boldsymbol{X}$. Uma matriz quadrada A é dita **unitária** se $A'A = AA' = \mathbf{I}$. Mostre que, se A é uma matriz $(d+1) \times (d+1)$ unitária, então $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}A) = A'\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$.



Exercício 35 (Paradoxo de Simpson)**.** Considere o gráfico de dispersão dos pares (y_i, x_{i1}), $i \in \{1, \ldots, n\}$ apresentado ao lado. Seja D_i definida, para $i \in \{1, \ldots, n\}$, pondo-se $D_i = 1$ se $x_{i1} > 2$ e $D_i = 0$ se $x_{i1} \leq 2$. Sejam ainda $\boldsymbol{X}$ e $\boldsymbol{Z}$ as matrizes definidas por

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix}, \qquad \boldsymbol{Z} = \begin{pmatrix} 1 & x_{1,1} & D_1 \\ 1 & x_{2,1} & D_2 \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & D_n \end{pmatrix}.$$

Usando a mesma notação do exercício 34, você consegue intuir, sem fazer as contas, qual será o sinal de $\widehat{\beta}_1(\boldsymbol{y}, \boldsymbol{X})$? E quanto a $\widehat{\beta}_1(\boldsymbol{y}, \boldsymbol{Z})$?

Código para gerar o gráfico do exercício no RStudio:

```
set.seed(3)
n=40
x = runif(n,0,4)
D = (x>2)
y = 2 - 4*D + x + rnorm(n, sd=.1)
plot(y~x, pch=16)
```

Esse exercício é um exemplo do *Paradoxo de Simpson*. Kadane[43] ilustra-o assim:

> "Imagine two routes to the summit of a mountain, a difficult route D and an easier route $\bar{D}$. Imagine also two groups of climbers: amateurs A, and experienced climbers, $\bar{A}$. Suppose that a person has probabilities of reaching the summit R, as a function of the route and the experience of the climber as follows:
>
> $$P\{R|\bar{D}, \bar{A}\} = 0.8 \quad P\{R|\bar{D}, A\} = 0.7$$
> $$P\{R|D, \bar{A}\} = 0.4 \quad P\{R|D, A\} = 0.3$$
>
> Thus experienced climbers are more likely to reach the summit whichever route they take, and both groups are less likely to reach the summit using the more difficult route.
>
> Further suppose also that the experienced climbers are believed to be more likely to take the more difficult route: $P\{D|\bar{A}\} = 0.75$ [and] $P\{D|A\} = 0.35$ ...
>
> ... The events $R\bar{D}$ and RD are disjoint, and their union is R. Therefore, $P\{R|\bar{A}\} = \cdots = 0.5$. Similarly $P\{R|A\} = \cdots = 0.56$. Thus amateur climbers have a greater chance of reaching the summit (0.56) than do experienced climbers (0.5), although for each route they have a smaller chance."

[43] Joseph B. Kadane. *Principles of Uncertainty*. Chapman and Hall/CRC, 2011

Propriedades amostrais do estimador de mínimos quadrados

> "Se sofres por algo externo a ti, o que te causa sofrimento não é a coisa em si mesma, mas sim a tua *estimativa* a respeito da coisa. Tens o poder de eliminar o sofrimento imediatamente."
>
> Marcus Aurelius Antoninus, Meditações, Livro VIII.

Estátua equestre de Marco Aurélio, Roma.

Após ficarem sumidos durante o capítulo pregresso, os símbolos $\mathbb{P}$ e $\mathbb{E}$ estão de volta! Quer dizer, uma vez mais estamos lidando com objetos (nossos velhos conhecidos $\boldsymbol{y}$, os $\boldsymbol{x}_i$'s, $\boldsymbol{X}$...) que são *aleatórios* — no sentido formal de que são funções cujo domínio é o espaço amostral $(\Omega, \mathscr{A})$ e cujo contradomínio é a reta, ou um dos espaços Euclidianos (do qual, de fato, a reta já é um exemplar e portanto não precisaria ter sido mencionada na abertura dessa lista), ou algum espaço de matrizes (que, já vimos, também é indistinguível,[44] do ponto de vista de sua estrutura linear e topológica, de um espaço Euclidiano). Significativamente, chamamos esses objetos de *aleatórios* antes mesmo de amuniciarmos o espaço amostral com uma medida de probabilidade $\mathbb{P}$, embora ao fim e ao cabo seja ela quem capture — ao transpormos o simbolismo para a linguagem natural — a ideia de que aqueles objetos são aleatórios. Essa digressão é relevante pois, no decorrer deste capítulo e dos seguintes, faremos suposições distribucionais — por exemplo, de exogeneidade estrita, de Gaussianidade dos termos de erros, de que estamos diante de um mecanismo de amostragem aleatória etc. — e é importante se dar conta que esses atributos distribucionais não estão inteiramente codificados nas variáveis, vetores e matrizes aleatórias com que lidamos, mas sim emergem como uma propriedade em que entram

[44] Vide o exercício 16

como ingredientes tanto essas variáveis, vetores, matrizes aleatórias *quanto também a medida de probabilidade* $\mathbb{P}$.[45]

[45] Veja o exercício 36.

Estimadores

Estimadores são variáveis (ou vetores, ou matrizes) aleatórias. Simples assim.[46] Um exemplo é o **estimador de mínimos quadrados ordinários**,[47] $\widehat{\boldsymbol{\beta}}$; o fato de que este é um vetor aleatório está implícito na expressão

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \equiv \left(\sum\nolimits_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \sum\nolimits_{i=1}^{n} \boldsymbol{x}_i y_i.$$

De fato, embora raramente recorramos a tão detalhada notação, podemos escrever[48]

$$\widehat{\boldsymbol{\beta}}(\omega) = \big(\boldsymbol{X}(\omega)'\boldsymbol{X}(\omega)\big)^{-1}\boldsymbol{X}(\omega)'\boldsymbol{y}(\omega), \qquad \omega \in \Omega,$$

em que — por óbvio — $[\boldsymbol{X}(\omega)]_{ij} = x_{ij}(\omega)$ etc. Vez que outra é conveniente fazer uma distinção, chamando de *estimador* ao vetor aleatório $\widehat{\boldsymbol{\beta}}$ e de uma *estimativa* ao valor realizado $\widehat{\boldsymbol{\beta}}(\omega) \in \mathbb{R}^{d+1}$. Essa é mesma distinção que fazemos ao chamar de *amostra aleatória* (no caso univariado) às variáveis aleatórias $y_1, \ldots, y_n$ e de uma *realização dessa amostra aleatória* aos números reais $y_1(\omega), \ldots, y_n(\omega)$.[49]

[46] Bom, não tão simples porque a variável aleatória deve estar definida "antes" de introduzirmos uma medida de probabilidade no espaço amostral; variáveis aleatórias do dipo $\vartheta = \mathbb{E}(x)$ "não valem"!

[47] Poderíamos tê-lo assim batizado já nos capítulos anteriores, mas optamos por não fazê-lo pois o epíteto 'estimador' em nada nos auxiliaria a compreensão.

[48] Anteriormente havíamos introduzido a notação $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$, cuja dependência em ω aqui apareceria, com um ligeiro abuso notacional, assim: $\widehat{\boldsymbol{\beta}}(\omega) = \widehat{\boldsymbol{\beta}}(\boldsymbol{y}(\omega), \boldsymbol{X}(\omega))$.

[49] Na nossa notação, escreveríamos $y_i(\omega) = y_i^{\text{obs}}$ etc.

O QUE NOS INTERESSA do ponto de vista de inferência estatística, contudo e sobretudo, é estabelecer alguma relação entre estimadores e *parâmetros*. Um **parâmetro** é um objeto da forma $\boldsymbol{\theta}_{\mathbb{P}} \in \Theta$, em que Θ é — tipicamente — um subconjunto de $\mathbb{R}^k$, onde $k \in \mathbb{N}$,[50] ou, mais geralmente, podemos enxergar o parâmetro como sendo a aplicação $\mathbb{P} \mapsto \boldsymbol{\theta}_{\mathbb{P}}$, no ímpeto de manter sempre na memória o fato de que *parâmetros dependem da medida de probabilidade* $\mathbb{P}$. Um exemplo de parâmetro é o vetor[51]

[50] E nada impede $\Theta = \mathbb{R}^k$!

$$\boldsymbol{\beta} = [\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})]^{-1}\mathbb{E}(\boldsymbol{X}'\boldsymbol{y}) \equiv \left(\sum\nolimits_{i=1}^{n} \mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')\right)^{-1} \sum\nolimits_{i=1}^{n} \mathbb{E}(\boldsymbol{x}_i y_i).$$

[51] A dependência de $\boldsymbol{\beta}$ em $\mathbb{P}$ está implícita, já que $\mathbb{E}$ depende de $\mathbb{P}$.

Em particular, se

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \tag{18}$$

é uma amostra iid, então vale a igualdade[52]

[52] Recorde o exercício 8!

$$\boldsymbol{\beta} = \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1). \tag{19}$$

Uma relação de particular interesse entre um estimador $\boldsymbol{\vartheta}\colon \Omega \to \mathbb{R}^k$ e um parâmetro $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_{\mathbb{P}} \in \mathbb{R}^k$ é o **viés**, definido por

$$\operatorname{bias}(\boldsymbol{\vartheta}, \boldsymbol{\theta}) \coloneqq \mathbb{E}(\boldsymbol{\vartheta}) - \boldsymbol{\theta}. \tag{20}$$

Um estimador $\boldsymbol{\vartheta}$ é dito **não-enviesado** para um parâmetro $\boldsymbol{\theta}$ se $\mathbb{E}(\boldsymbol{\vartheta}) = \boldsymbol{\theta}$, isto é, se $\operatorname{bias}(\boldsymbol{\vartheta}, \boldsymbol{\theta}) = \mathbf{0}$.[53]

[53] Note que a expressão definindo o viés depende da medida de probabilidade $\mathbb{P}$. Em particular — e isso não é lá muito óbvio — um estimador pode ser não enviesado para um parâmetro $\boldsymbol{\theta}$ sob uma medida de probabilidade e ser enviesado sob outra. Veja o exercício 37.

De posse desse vocabulário, podemos agora nos perguntar se o estimador de MQO é não-enviesado para o $\boldsymbol{\beta}$ populacional. A resposta é que, em geral, $\operatorname{bias}\big(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\big) \neq \mathbf{0}$; o seguinte resultado oferece, como corolário, condições suficientes e necessárias para que esse viés seja nulo.

Teorema 17. Sejam $\boldsymbol{y}$ e $\boldsymbol{x}_i$, $i \in \{1, \dots, n\}$, vetores aleatórios em $\mathbb{R}^n$ e $\mathbb{R}^{d+1}$, respectivamente, e seja $\boldsymbol{X}$ a matriz aleatória $n \times (d+1)$ cuja componente i, j é x_{ij}. Seja ainda $\mathbb{P}$ uma medida de probabilidade em $(\Omega, \mathscr{A})$ satisfazendo

(i) $\mathbb{P}\big(\boldsymbol{\xi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\xi} > 0 \text{ para todo } \boldsymbol{\xi} \in \mathbb{R}^{d+1} \text{ não nulo}\big) = 1$.

(ii) $\mathbb{E}_{\mathbb{P}}\big|\widehat{\boldsymbol{\beta}}\big| < \infty$, onde $\widehat{\boldsymbol{\beta}} \coloneqq (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

Então,

$$\mathbb{E}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \boldsymbol{X}\big) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{u} \,|\, \boldsymbol{X}), \tag{21}$$

em que $\boldsymbol{\beta} \coloneqq [\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})]^{-1}\mathbb{E}(\boldsymbol{X}'\boldsymbol{y}) \in \mathbb{R}^{d+1}$ e onde $\boldsymbol{u}$ é o vetor aleatório em $\mathbb{R}^n$ definido por $\boldsymbol{u} \coloneqq \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.

Corolário 18. Nas condições do Teorema 17, tem-se $\operatorname{bias}\big(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\big) = \mathbf{0}$ se, e somente se, $\mathbb{E}\big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\big) = \mathbf{0}$. Em particular, se $\boldsymbol{X}$ é estritamente exógeno, então $\widehat{\boldsymbol{\beta}}$ é não-enviesado para $\boldsymbol{\beta}$.

Demonstração. Primeiramente, verifiquemos que o estimador $\widehat{\boldsymbol{\beta}}$ e o parâmetro $\boldsymbol{\beta}$ estão bem definidos. Quanto a $\widehat{\boldsymbol{\beta}}$, é imediato, pois a condição (i) impõe nada mais que a invertibilidade da matriz $\boldsymbol{X}(\omega)'\boldsymbol{X}(\omega)$, para ω pertencente a um evento $E^* \subseteq \Omega$ com $\mathbb{P}(E^*) = 1$. Quanto a $\boldsymbol{\beta}$, vemos que, para qualquer $\boldsymbol{\xi} \in \mathbb{R}^{d+1}$, vale a desigualdade $\mathbb{E}(\boldsymbol{\xi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\xi}) > 0$, novamente pela condição (i). Disso segue, usando a identidade $\boldsymbol{\xi}'\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\xi} = \mathbb{E}(\boldsymbol{\xi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\xi})$, que a matriz $\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})$ é invertível.

Para a tese temos, das definições, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$. Logo,

$$\widehat{\boldsymbol{\beta}} \coloneqq (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}\big) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}.$$

Tomando a esperança incondicional em ambos os lados dessa identidade, obtemos a caracterização declarada no corolário. Tomando a esperança condicional em $\boldsymbol{X}$, seguem o teorema e a condição suficiente para anular o viés no corolário. ■

Quanto à dispersão de $\widehat{\boldsymbol{\beta}}$, temos o seguinte.

Teorema 19. Nas condições do Teorema 17, seja

$$\boldsymbol{X}_* \coloneqq (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

Então

1. $\mathbb{V}\big(\widehat{\boldsymbol{\beta}}\big) = \mathbb{E}\big(\boldsymbol{X}_*\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}_*'\big) - \mathbb{E}\big(\boldsymbol{X}_*\boldsymbol{u}\big)\mathbb{E}\big(\boldsymbol{u}'\boldsymbol{X}_*'\big)$.
2. $\mathbb{V}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \boldsymbol{X}\big) = \boldsymbol{X}_*\mathbb{V}\big(\boldsymbol{u} \,\big|\, \boldsymbol{X}\big)\boldsymbol{X}_*'$

Adicionalmente,

3. se $\operatorname{bias}\big(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}\big) = 0$, então
$$\mathbb{V}\big(\widehat{\boldsymbol{\beta}}\big) = \mathbb{E}\Big\{\mathbb{V}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \boldsymbol{X}\big)\Big\} + \mathbb{E}\Big\{\boldsymbol{X}_*\mathbb{E}(\boldsymbol{u} \,|\, \boldsymbol{X})\mathbb{E}(\boldsymbol{u}' \,|, \boldsymbol{X})\boldsymbol{X}_*'\Big\}.$$
4. se $\boldsymbol{X}$ é estritamente exógeno, então
$$\mathbb{V}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \boldsymbol{X}\big) = \boldsymbol{X}_*\mathbb{E}(\boldsymbol{u}\boldsymbol{u}' \,|\, \boldsymbol{X})\boldsymbol{X}_*'.$$
5. se os termos de erro são **homoscedásticos**, no sentido de que $\mathbb{V}\big(\widehat{\boldsymbol{u}} \,\big|\, \boldsymbol{X}\big) = \sigma^2\mathbf{I}$ para alguma constante $\sigma > 0$, então $\mathbb{V}\big(\widehat{\boldsymbol{\beta}} \,\big|\, \boldsymbol{X}\big) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

O Teorema de Gauss–Markov

O Teorema de Gauss–Markov é a cereja do bolo no estudo das propriedades estatísticas do estimador de MQO em amostras finitas. Ele afirma que, sob certas condições, o estimador $\widehat{\boldsymbol{\beta}}$ possui a propriedade de ser o *melhor estimador linear não-enviesado* para $\boldsymbol{\beta}$ — uma propriedade que, em inglês, recebe a carinhosa alcunha de BLUE (*best linear unbiased estimator*). Em que pese sua importância, esse resultado — tal qual uma cereja não figurativa em um bolo não figurativo — deve ser usado com ponderação. Não devemos nos empolgar com a aparente invencibilidade do nosso amigo "beta-chapéu". Por três razões, pelo menos: primeiro, uma das hipóteses

do teorema é de que os termos de erro na equação $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ são homoscedásticos, isto é satisfazem a condição $\mathbb{V}\big(\widehat{\boldsymbol{u}} \,\big|\, \boldsymbol{X}\big) = \sigma^2\mathbf{I}$. Essa hipótese é, em um sem fim de casos, inadequada, pois impede que a variância da resposta dependa do nível dos regressores, algo que raramente ocorre em dados observacionais. Segundo, sua tese restringe-se à classe de estimadores *lineares* e *não-enviesados* e, sem a ressalva de que podem existir estimadores *não-lineares* ou *enviesados* para $\boldsymbol{\beta}$ que sejam "melhores" do que $\widehat{\boldsymbol{\beta}}$, o teorema soa mais poderoso do que realmente é. Terceiro, a noção de melhor aqui refere-se à ordem parcial no espaço das matrizes de covariância $(d+1) \times (d+1)$, em que uma matriz A é dita "maior" do que uma matriz B sse $A - B$ é positiva semidefinida; nada impede, todavia, que avaliemos a qualidade de um estimador segundo outros critérios.

Teorema 20 (Gauss–Markov)**.** Nas condições do Teorema 17, suponha adicionalmente que

1. $\mathbb{E}(\boldsymbol{u} \,|\, \boldsymbol{X}) = \mathbf{0}$ quase certamente. (exogeneidade)
2. $\mathbb{E}(\boldsymbol{u}\boldsymbol{u}' \,|\, \boldsymbol{X}) = \sigma^2\mathbf{I}$, para algum $\sigma > 0$. (homoscedasticidade)

Seja ainda $\boldsymbol{A}$ uma matriz $n \times (d+1)$ da forma $\boldsymbol{A} = \varphi \circ \boldsymbol{X}$, onde φ é uma aplicação de $\mathbb{R}^{n\times(d+1)}$ em $\mathbb{R}^{n\times(d+1)}$. Se $\mathbb{E}(\boldsymbol{A}'\boldsymbol{y} \,|\, \boldsymbol{X}) \overset{q.c.}{=} \boldsymbol{\beta}$, então $\mathbb{V}(\boldsymbol{A}'\boldsymbol{y} \,|\, \boldsymbol{X}) - \mathbb{V}(\widehat{\boldsymbol{\beta}} \,|\, \boldsymbol{X})$ é positiva semidefinida.

Exercícios

Quaisquer erros nos enunciados são — provavelmente! — intencionais.

Exercício 36. Considere o espaço amostral

$$\big(\Omega, \mathscr{A}\big) \coloneqq \big([0,1], \text{"Borelianos em } [0,1]\text{"}\big)$$

e a variável aleatória definida, para $\omega \in \Omega$, por $x(\omega) \coloneqq \omega$. Sejam $\mathbb{P}_1$ e $\mathbb{P}_2$ as medidas de probabilidade em $(\Omega, \mathscr{A})$ definidas, respectivamente, por

$$\mathbb{P}_1[a,b] = b - a, \qquad 0 \le a < b \le 1$$

e

$$\mathbb{P}_2[a,b] = \frac{1}{2}\mathbb{I}_{\{0\in[a,b]\}} + \frac{1}{2}\mathbb{I}_{\{1\in[a,b]\}}, \qquad 0 \le a < b \le 1.$$

Verifique que $x \sim \text{Uniforme}[0,1]$ no espaço de probabilidade $(\Omega, \mathscr{A}, \mathbb{P}_1)$ e que $x \sim \text{Bernoulli}(1/2)$ no espaço de probabilidade $(\Omega, \mathscr{A}, \mathbb{P}_2)$.

Exercício 37. Nas mesmas condições do exercício acima, seja $\mathbb{P}_3$ a medida de probabilidade definida por

$$\mathbb{P}_3[a,b] = b^2 - a^2, \qquad 0 \leq a < b \leq 1.$$

Considere o parâmetro $\theta_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}(x)$, onde o subscrito na esperança indica a medida de probabilidade subjacente. Considere ainda os estimadores ϑ_1 e ϑ_2 definidos, para $\omega \in \Omega$, respectivamente por

$$\vartheta_1(\omega) \coloneqq x(\omega) \quad \text{e} \quad \vartheta_2(\omega) \coloneqq \frac{1}{2}.$$

1. Mostre que ϑ_1 é não-enviesado para $\theta_{\mathbb{P}}$ sob qualquer uma das medidas de probabilidade $\mathbb{P}_1$, $\mathbb{P}_2$ e $\mathbb{P}_3$.

2. Mostre que ϑ_2 é não-enviesado para $\theta_{\mathbb{P}}$ sob as medidas de probabilidade $\mathbb{P}_1$ e $\mathbb{P}_2$, mas é enviesado para $\theta_{\mathbb{P}_3}$.

Exercício 38. Mostre que $\mathbb{V}(\boldsymbol{z}) = \mathbb{E}\big(\mathbb{V}(\boldsymbol{z} \,|\, \boldsymbol{X})\big) + \mathbb{V}\big(\mathbb{E}(\boldsymbol{z} \,|\, \boldsymbol{X})\big)$.

Exercício 39. Nas condições do Teorema 17, mostre que

$$\operatorname{cov}\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}} \,|\, \boldsymbol{X}\right) = \mathbf{0}.$$

Exercício 40. Demonstre o Teorema 19.

Exercício 41. Mostre que, nas condições do Teorema 17, se $\boldsymbol{u}$ e $\boldsymbol{X}$ são independentes, com $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, então

$$\widehat{\boldsymbol{\beta}} \,|\, \boldsymbol{X} \sim \mathrm{N}\big(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\big).$$

Conclua que a distribuição incondicional de $\widehat{\boldsymbol{\beta}}$ é uma **mistura** de distribuições Normais.

Exercício 42. Nas condições do Teorema 17, considere o estimador

$$\widehat{\sigma}^2 \coloneqq \frac{|\widehat{\boldsymbol{u}}|^2}{n - (d+1)}.$$

Mostre que, se $\boldsymbol{X}$ é estritamente exógeno e $\mathbb{V}(\boldsymbol{u} \,|\, \boldsymbol{X}) = \sigma^2\mathbf{I}$, para algum $\sigma > 0$, então $\mathbb{E}\big(\widehat{\sigma}^2\big) = \sigma^2$. A raiz $\widehat{\sigma}$ é dita o **erro padrão da regressão**.

Exercício 43. Nas condições do Teorema 17, seja Π a matriz de projeção ortogonal sobre $\operatorname{col}(\boldsymbol{X})$, isto é, $\Pi \coloneqq \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, e seja $\Pi^{\perp} \coloneqq \mathbf{I} - \Pi$ a projeção ortogonal sobre $\operatorname{col}(\boldsymbol{X})^{\perp}$. Mostre que $\widehat{\boldsymbol{u}} = \Pi^{\perp}\boldsymbol{u}$. Conclua que $|\widehat{\boldsymbol{u}}|^2 = \boldsymbol{u}'\Pi^{\perp}\boldsymbol{u}$.

Exercício 44 (Medidas empíricas)**.** Seja $n \in \mathbb{N}$ fixado. Suponha que

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \tag{22}$$

é uma amostra aleatória de $(\boldsymbol{x}, y)$. Para cada $\omega \in \Omega$, seja $\widehat{\mathbb{P}}_{n,\omega}$ a medida de probabilidade (discreta) dada por

$$\widehat{\mathbb{P}}_{n,\omega}\big[(\boldsymbol{x}, y) = \big(\boldsymbol{x}_i(\omega), y_i(\omega)\big)\big] = \frac{1}{n}, \quad \forall i \in \{1, \ldots, n\}.$$

Isto é, $\widehat{\mathbb{P}}_n$ é a medida de probabilidade (aleatória) em Ω que atribui massa $1/n$ para cada um dos n pontos da amostra (22).[54] Escreva $\widehat{\mathbb{E}}_{n,\omega}$ para denotar o valor esperado com respeito à medida de probabilidade $\widehat{\mathbb{P}}_{n,\omega}$. Mostre que, para qualquer $\varphi\colon \mathbb{R}^{d+1} \times \mathbb{R} \to \mathbb{R}^k$ contínua, vale

$$\widehat{\mathbb{E}}_{n,\omega}\big(\varphi \circ (\boldsymbol{x}, y)\big) = \frac{1}{n}\sum_{i=1}^{n} \varphi\big(\boldsymbol{x}_i(\omega), y_i(\omega)\big), \quad \omega \in \Omega.$$

Conclua que

$$\widehat{\boldsymbol{\beta}}(\omega) = \Big(\widehat{\mathbb{E}}_{n,\omega}(\boldsymbol{x}_1\boldsymbol{x}_1')\Big)^{-1}\widehat{\mathbb{E}}_{n,\omega}(\boldsymbol{x}_1 y_1)$$

e compare com a equação (19).

Exercício 45. No contexto de um modelo linear com regressores estocásticos, que é o caso em quase toda a literatura de econometria, o Teorema de Gauss–Markov costuma ser enunciado de maneira não completamente rigorosa, algo que só se percebe ao ler cuidadosamente a sua demonstração. Em particular, o termo "não-enviesado" é empregado sem deixar claro que, na prova, será usada a propriedade mais forte de que o estimador deve ser *condicionalmente não-enviesado.* De fato, se considerarmos estimadores incondicionalmente não-enviesados, o Teorema de Gauss–Markov pode deixar de valer.[55] Aqui está um enunciado rigoroso do teorema (o exercício é demonstrá-lo):

[54] $\widehat{\mathbb{P}}_n$ é dita uma *medida empírica.* Um detalhe é que essa medida de probabilidade, em geral, não pode ser definida na σ-álgebra original $\mathscr{A}$, mas somente naquela induzida pela amostra aleatória, isto é, em $\sigma(\boldsymbol{y}, \boldsymbol{X})$.

[55] Juliet Popper Shaffer. The Gauss-Markov Theorem and random regressors. *The American Statistician*, 45(4):269–273, 1991

Teorema 21. Considere dados um espaço amostral $(\Omega, \mathscr{A})$, uma matriz $n \times d$ aleatória, $\boldsymbol{X}$, e um vetor $n \times 1$ aleatório, $\boldsymbol{v}$. Seja $\mathfrak{M}$ a coleção de todas as medidas de probabilidade $\mathbb{P}\colon \mathscr{A} \to \mathbb{R}$ que satisfazem o seguinte:

1. $\mathbb{P}(\operatorname{rank}(\boldsymbol{X}) = d) = 1$
2. $\mathbb{E}_{\mathbb{P}}\big(\boldsymbol{v}'\boldsymbol{v}\big) < \infty$
3. $\mathbb{E}_{\mathbb{P}}(\boldsymbol{v}|\boldsymbol{X}) = \mathbf{0}$
4. $\mathbb{E}_{\mathbb{P}}(\boldsymbol{v}\boldsymbol{v}'|\boldsymbol{X}) = \mathbf{Id}$

Seja ainda $\psi : M(n \times d) \to M(n \times d)$ uma aplicação mensurável (onde $M(n \times d)$ denota o espaço vetorial das matrizes $n \times d$), e ponha $\boldsymbol{X}_\psi = \psi \circ \boldsymbol{X}$.

Suponha que, para toda $\mathbb{P} \in \mathfrak{M}$, todo $\boldsymbol{\beta} \in \mathbb{R}^d$ e todo $\sigma > 0$, vale

$$\mathbb{E}_{\mathbb{P}}(\boldsymbol{X}'_\psi(\boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{v}) \,|\, \boldsymbol{X}) = \boldsymbol{\beta}, \qquad \mathbb{P}\text{-q.c.}$$

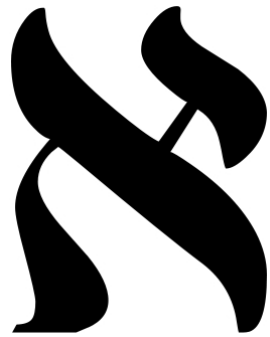
Então, para qualquer $\mathbb{P} \in \mathfrak{M}$, qualquer $\boldsymbol{\beta} \in \mathbb{R}^d$ e qualquer $\sigma > 0$, tem-se que a matriz $d \times d$

$$\mathbb{V}_{\mathbb{P}}(\boldsymbol{X}'_\psi(\boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{v}) \,|\, \boldsymbol{X}) - \mathbb{V}_{\mathbb{P}}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{v}) \,|\, \boldsymbol{X})$$

é positiva semidefinida.

περὶ τῶν ἀσυμπτώτων *(sobre assíntotas)*

> "En la parte inferior del escalón, hacia la derecha, vi una pequeña esfera tornasolada, de casi intolerable fulgor. Al principio la creí giratoria; luego comprendí que ese movimiento era una ilusión producida por los vertiginosos espectáculos que encerraba. El diámetro del Aleph sería de dos o tres centímetros, pero el espacio cósmico estaba ahí, sin disminución de tamaño."
>
> J. L. Borges, *El Aleph*.



O conto *El Aleph*, de Borges, inicia com uma citação da famosa passagem em que Hamlet considera a possibilidade de viver recluso numa casca de noz e, ainda assim, proclamar-se Rei do espaço infinito. Ilustrado nesse trecho está o truísmo de que não compreendemos, intuitivamente, o conceito de infinito — ou ao menos o fato de que esse conceito pode ser enganoso! Pois bem, iniciaremos neste capítulo uma jornada, que será cumprida no próximo, cuja meta é estudar as propriedades distribucionais do estimador de mínimos quadrados ordinários *quando o tamanho da amostra vai para o infinito*.[56] Ocorre que amostras são sempre finitas: estamos sempre confinados ao interior, às fronteiras, duma casca de um fruto da nogueira. Por essa razão, faz sentido ter em mente que aquilo que estaremos estudando são, na verdade, aproximações as quais são válidas em amostras "suficientemente grandes" (mas ainda finitas). As duas mais importantes dessas propriedades — pertinentes ao estimador de MQO, sob certas suposições — são sua *consistência* e *normalidade assintóticas*. O primeiro passo será introduzir e estudar os conceitos de convergência a partir dos quais serão tornadas precisas as noções de consistência e normalidade assintótica mencionadas acima. Trataremos do estimador de mínimos quadrados no capítulo seguinte.

[56] A essas costumamos chamar de *propriedades assintóticas*.

Convergência em probabilidade

Suponha que nos é dada uma sequência $(\boldsymbol{z}_n)_{n\geq 1} \equiv (\boldsymbol{z}_1, \boldsymbol{z}_2, \dots)$ de vetores aleatórios em $\mathbb{R}^k$, onde $k \in \mathbb{N}$ está fixado, e seja $\boldsymbol{z}$ também um vetor aleatório em $\mathbb{R}^k$. Dizemos[57] que a sequência $(\boldsymbol{z}_n)_{n\geq 1}$ **converge em probabilidade para** $\boldsymbol{z}$ se, e somente se, valer a seguinte propriedade: dado $\delta > 0$ arbitrário, é possível encontrar um número natural n_δ tal que valha a desigualdade

$$\mathbb{P}\big(\big|\boldsymbol{z}_n - \boldsymbol{z}\big| > \delta\big) < \delta \quad \text{para todo } n \geq n_\delta.$$

Isso é o mesmo que dizer que $\lim_{n\to\infty} \mathbb{P}\big(\big|\boldsymbol{z}_n - \boldsymbol{z}\big| > \delta\big) = 0$, para qualquer $\delta > 0$.[58] Nessas condições, dizemos também que $\boldsymbol{z}$ é o **limite em probabilidade** da sequência $(\boldsymbol{z}_n)_{n\geq 1}$, e escrevemos

(i) $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{z}$;
(ii) $\operatorname{plim}_{n\to\infty} \boldsymbol{z}_n = \boldsymbol{z}$;
(iii) "$\boldsymbol{z}_n \to \boldsymbol{z}$ em probabilidade",

etc. Se $\boldsymbol{z}$ é uma variável aleatória degenerada (digamos, com $\mathbb{P}(\boldsymbol{z} = \boldsymbol{\xi}) = 1$ para algum $\boldsymbol{\xi} \in \mathbb{R}^k$) então dizer que $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{\xi}$ é o mesmo que afirmar que, para valores de n "suficientemente grandes", a probabilidade (ou, para sermos bem precisos, a $\mathbb{P}$–probabilidade) de encontrarmos a variável aleatória $\boldsymbol{z}_n$ na região

$$\overline{\text{ball}}(\boldsymbol{\xi}; \delta) \coloneqq \{\boldsymbol{\eta} \in \mathbb{R}^k \colon\ |\boldsymbol{\eta} - \boldsymbol{\xi}| \leq \delta\}$$

pode ser tomada tão próxima de 1 quanto queiramos (enfatizando: desde que o índice n seja tomado suficientemente grande).

Exemplo 22. No espaço de probabilidade do exercício 22, considere a sequência de variáveis aleatórias $(z_n)_{n\geq 1}$ definidas, para $\omega \in (0, 1]$ e $n \geq 1$, via

$$z_n(\omega) \coloneqq \begin{cases} 1, & \text{se } n \leq (1+\omega)\cdot 2^{f(n)} < n+1, \\ 0, & \text{caso contrário,} \end{cases}$$

onde $f(n) \coloneqq \lfloor \log_2(n) \rfloor$ = "maior número inteiro majorado por $\log_2(n)$". Vejamos que $z_n \xrightarrow{p} 0$. Seja $\delta > 0$. Se $\delta \geq 1$ podemos tomar $n_\delta = 1$, pois nesse caso de fato temos, para todo $n \geq 1$,

$$\mathbb{P}(|z_n - 0| > \delta) = \mathbb{P}(\varnothing) = 0 < \delta,$$

[57] Se quisermos ser extremamente precisos na terminologia, devemos perceber que a medida de probabilidade $\mathbb{P}$ está aí implícita: poderíamos com maior exatidão dizer que $(\boldsymbol{z}_n)_{n\geq 1}$ converge em $\mathbb{P}$–probabilidade para $\boldsymbol{z}$.

[58] Veja o exercício 46.

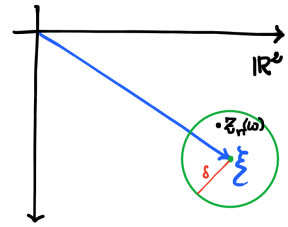


Figura 4: A figura exemplifica um ω pertencente ao evento $\{|\boldsymbol{z}_n - \boldsymbol{\xi}| \leq \delta\}$, o qual poderíamos reescrever como $\{\boldsymbol{z}_n \in \overline{\text{ball}}(\boldsymbol{\xi}; \delta)\}$. Se $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{\xi}$, o conjunto de tais ω's tem probabilidade próxima de 1, para todo índice n a partir de n_δ.

já que $z_n(\cdot) \leq 1$ para todo n. Por outro lado, se $0 < \delta < 1$, então temos a igualdade de eventos

$$\big[|z_n - 0| > \delta\big] = \big[z_n = 1\big] = \big\{\omega : n \leq (1+\omega)\cdot 2^{f(n)} < n+1\big\}.$$

Quer dizer, para $\delta \in (0,1)$ e $n \geq 1$ o evento $\big[|z_n - 0| > \delta\big]$ coincide com o intervalo

$$\left(\frac{n}{2^{f(n)}} - 1,\ \frac{n+1}{2^{f(n)}} - 1\right].$$

Daí, pela definição de $\mathbb{P}$, temos $\mathbb{P}(|z_n - 0| > \delta) = 2^{-f(n)}$. Escrevendo $f(n) = \log_2(n) - h(n)$, onde $h(n) \in [0,1)$, temos $2^{-f(n)} = 2^{h(n)}/n \leq 2/n$ e, tomando $n_\delta := \lceil 4/\delta \rceil$, obtemos finalmente $\mathbb{P}(|z_n - 0| > \delta) < \delta$ para todo $n \geq n_\delta$.

Convergência em distribuição

Antes de definirmos convergência em distribuição, precisamos de dois lembretes. O primeiro deles, de que a **função de distribuição acumulada**[59] de um vetor aleatório $\boldsymbol{z}$ em $\mathbb{R}^k$ é definida como sendo a função $F_{\boldsymbol{z}} \colon \mathbb{R}^k \to \mathbb{R}$ cuja regra é

$$F_{\boldsymbol{z}}(\boldsymbol{\xi}) := \mathbb{P}(z_1 \leq \xi_1, \ldots, z_k \leq \xi_k), \quad \boldsymbol{\xi} \equiv (\xi_1\ \cdots\ \xi_k)' \in \mathbb{R}^k.$$

O segundo, de que uma função $h \colon \mathbb{R}^k \to \mathbb{R}^d$ é dita **contínua no ponto** $\boldsymbol{\xi} \in \mathbb{R}^k$ se, e somente se, vale a implicação

$$\boldsymbol{\xi}_n \to \boldsymbol{\xi} \in \mathbb{R}^k \quad \Longrightarrow \quad h(\boldsymbol{\xi}_n) \to h(\boldsymbol{\xi}) \in \mathbb{R}^d.$$

Se h é contínua no ponto $\boldsymbol{\xi}$, chamemo-lo então – mui apropriadamente – de um **ponto de continuidade de** h. Caso h seja contínua em todos os pontos do seu domínio, dizemos que h é **contínua**.[60]

Sejam $(\boldsymbol{z}_n)_{n\geq 1}$ e $\boldsymbol{z}$ como na seção anterior. Podemos, enfim, definir que a sequência $(\boldsymbol{z}_n)_{n\geq 1}$ **converge em distribuição** para $\boldsymbol{z}$ se, e somente se, para cada ponto de continuidade $\boldsymbol{\xi} \in \mathbb{R}^k$ da função de distribuição acumulada $F_{\boldsymbol{z}}$, a sequência numérica $F_{\boldsymbol{z}_n}(\boldsymbol{\xi})$ convergir para $F_{\boldsymbol{z}}(\boldsymbol{\xi})$, isto é, se, e somente se, valer o seguinte:

"se $\boldsymbol{\xi}$ é ponto de contínuidade de $F_{\boldsymbol{z}}$, então $\lim_{n\to\infty} F_{\boldsymbol{z}_n}(\boldsymbol{\xi}) = F_{\boldsymbol{z}}(\boldsymbol{\xi})$".

Nesse contexto, dizemos que $\boldsymbol{z}$ é o **limite em distribuição** da sequência $(\boldsymbol{z}_n)_{n\geq 1}$, e escrevemos, e.g.

[59] A importância dessa noção é que uma medida de probabilidade em $\mathbb{R}^k$, $k \geq 1$, é unicamente determinadas por sua função de distribuição acumulada (e reciprocamente).

[60] Uma função h é contínua nesse sentido se, e somente se, $h^{-1}(U)$ é um subconjunto aberto de $\mathbb{R}^k$, para todo subconjunto U aberto em $\mathbb{R}^d$.

(a) $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$;
(b) $\boldsymbol{z}_n \xrightarrow{\mathcal{L}} \boldsymbol{z}$;
(c) $\mathcal{L}(\boldsymbol{z}_n) \to \mathcal{L}(\boldsymbol{z})$; (o $\mathcal{L}$ é para convergência em $\mathcal{L}$ei)
(d) "$\boldsymbol{z}_n \to \boldsymbol{z}$ em distribuição",

etc.

Na literatura de teoria da probabilidade, convergência em distribuição também é chamada de *convergência em lei*, de *convergência fraca*, e de *convergência fraca** (lê-se "convergência-fraca-estrela" mesmo), sendo essa última terminologia — proveniente da análise funcional — a mais adequada. É importante notar que convergência em distribuição realmente não diz respeito aos vetores aleatórios em si, mas sim às suas respectivas funções de distribuição acumuladas, e nesse sentido o vetor aleatório limítrofe, $\boldsymbol{z}$, aparece apenas de maneira auxiliar (de fato, até mesmo a sequência $(\boldsymbol{z}_n)_{n\geq 1}$ é dispensável, mas vamos com calma): se H é uma função de distribuição acumulada qualquer (com domínio $\mathbb{R}^k$) tal que $F_{\boldsymbol{z}_n}(\boldsymbol{\xi}) \to H(\boldsymbol{\xi})$ para qualquer $\boldsymbol{\xi}$ que seja ponto de continuidade de H, então dizemos que $(\boldsymbol{z}_n)_{n\geq 1}$ **converge em distribuição para** H, e escrevemos

(e) $\boldsymbol{z}_n \xrightarrow{d} H$.

Apropriadamente, a distribuição H é dita a **distribuição limite** da sequência $(\boldsymbol{z}_n)_{n\geq 1}$. Adicionalmente, se H possui um "código padrão" (por exemplo,[61] $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ou Uniforme$(0, 1]$ etc), escrevemos também, e.g.

(f) $\boldsymbol{z}_n \xrightarrow{d} \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Nesse último caso, dizemos que a sequência $(\boldsymbol{z}_n)_{n\geq 1}$ é **assintoticamente normal** (com *vetor de médias* $\boldsymbol{\mu}$ e *matriz de covariâncias* $\boldsymbol{\Sigma}$) e, no caso particular em que $\boldsymbol{\mu} = \mathbf{0}$ e $\boldsymbol{\Sigma} = \mathbf{Id}$, que a sequência é **assintoticamente normal padrão**. O conceito de convergência em distribuição é particularmente importante pois permite que calculemos probabilidades aproximadas quando a distribuição exata da variável aleatória $\boldsymbol{z}_n$ é desconhecida e difícil de se obter analiticamente. Quer dizer, se $\boldsymbol{z}_n \xrightarrow{d} H$, então podemos usar a aproximação $\mathbb{P}(\boldsymbol{z}_n \leq \boldsymbol{\xi}) \approx H(\boldsymbol{\xi})$, $\boldsymbol{\xi} \in \mathbb{R}^k$:

Proposição 23. Se $\boldsymbol{z}_n \xrightarrow{d} H$ e H é contínua, então para qualquer $\delta > 0$ é possível encontrar um número natural n_δ tal que a

[61] Lembre: uma variável aleatória z é dita ter **distribuição Normal (ou Gaussiana)** se z é degenerada ou se z possui função densidade de probabilidade dada, para $\xi \in \mathbb{R}$, por

$$f_z(\xi) = \frac{1}{s\sqrt{2\pi}} \mathrm{e}^{(\xi-m)^2/(2s^2)}$$

para algum par de parâmetros $m \in \mathbb{R}$ e $s > 0$. Um vetor aleatório $\boldsymbol{z}$ em $\mathbb{R}^k$ é dito ter **distribuição Normal multivariada** se, para todo $\boldsymbol{\xi} \in \mathbb{R}^k$, a variável aleatória $\boldsymbol{\xi}'\boldsymbol{z}$ tem distribuição Normal.

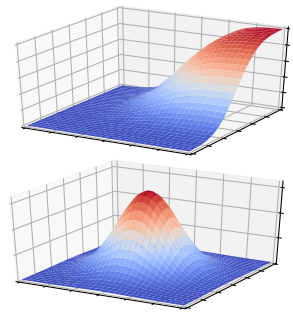


Figura 5: A função de distribuição acumulada de uma $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (topo) e a respectiva função densidade de probabilidade (abaixo).

desigualdade

$$\sup_{\boldsymbol{\xi}\in\mathbb{R}^d} \left|\mathbb{P}(\boldsymbol{z}_n \leq \boldsymbol{\xi}) - H(\boldsymbol{\xi})\right| < \delta$$

vale para todo $n \geq n_\delta$.

Nesse ponto é importante ressaltar que o espaço amostral $(\Omega, \mathscr{A})$ que estamos considerando deve comportar tais sequências infinitas de variáveis aleatórias. Não é trivial que um tal espaço exista; o *Teorema de Existência de Kolmogorov* nos garante que não temos com o que nos preocupar — de fato, o teorema é construtivo: ver Billingsley[62] para mais detalhes. Para nós, é suficiente enunciar a seguinte versão desse resultado:

[62] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 3rd edition, 1995

Teorema 24. Seja $(H_n)_{n\geq 1}$ uma sequência de funções de distribuição acumuladas com

$$H_n\colon \left(\mathbb{R}^k\right)^n \to \mathbb{R}, \quad n \geq 1.$$

Suponha que o critério de compatibilidade

$$H_{n+1}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n, +\infty) = H_n(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$$

é satisfeito para toda n–upla $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ de vetores em $\mathbb{R}^k$. Então existe um espaço de probabilidade $(\Omega, \mathscr{A}, \mathbb{P})$ e uma sequência $(\boldsymbol{z}_n)_{n\geq 1}$ de vetores aleatórios em $\mathbb{R}^k$ tais que, para todo $n \geq 1$ e toda n–upla $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ de vetores em $\mathbb{R}^k$, vale a igualdade

$$\mathbb{P}(\boldsymbol{z}_1 \leq \boldsymbol{\xi}_1, \ldots, \boldsymbol{z}_n \leq \boldsymbol{\xi}_n) = H_n(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n).$$

Uma importante aplicação do teorema acima é a seguinte: tomando uma função de distribuição acumulada fixa $H\colon \mathbb{R}^k \to \mathbb{R}$ e a partir dela definindo a sequência $(H_n)_{n\geq 1}$ via

$$H_n(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n) \coloneqq H(\boldsymbol{\xi}_1) \times \cdots \times H(\boldsymbol{\xi}_n),$$

para $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n \in \mathbb{R}^k$ e $n \geq 1$, o teorema nos garante que existe uma *sequência iid*[63] $(\boldsymbol{z}_n)_{n\geq 1}$ de vetores aleatórios em $\mathbb{R}^k$ com $\boldsymbol{z}_n \sim H$ para todo n. Quando fazemos teoria assintótica para dados *cross section*, embora na maior parte do tempo não tenhamos (nem queiramos ter) em mente qual é o espaço amostral sendo utilizado, necessitamos ao menos garantir que os objetos matemáticos com os quais estamos trabalhando sejam bem definidos — a presente aplicação do teorema de Kolmogorov supre essa necessidade.

[63] Uma sequência infinita de vetores aleatórios $(\boldsymbol{z}_n)_{n\geq 1}$ é iid se, e somente se, para cada $n \geq 1$ os vetores aleatórios $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ são iid, no sentido que vimos anteriormente.

Elementos de teoria assintótica

No próximo capítulo, estudaremos as propriedades assintóticas do estimador de mínimos quadrados ordinários. Veremos que, sob certas condições, vale que $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}$ e que $\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) \xrightarrow{d} \mathrm{N}\big(0, \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\big)$, onde a matriz de covariâncias $\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}$ será especificada oportunamente. Note que escrevemos $\widehat{\boldsymbol{\beta}}_n$ ao invés de simplesmente $\widehat{\boldsymbol{\beta}}$ para enfatizar que "o" estimador de MQO é na verdade um termo em uma sequência de estimadores: de fato, na fórmula $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ está implícita a dependência no tamanho da amostra, através das dimensões das matrizes e vetores que aparecem nessa expressão. Veja: com as definições dadas nas duas seções anteriores, já é possível compreender as propriedades de consistência e normalidade assintótica do estimador de MQO! Para provar essas propriedades, entretanto, vamos recorrer a uma dose adicional de teoria. Enunciaremos os principais resultados que serão usados posteriormente, sem prová-los. Algumas dessas demonstrações estão espalhadas nos exercícios logo adiante; os detalhes estão no *Asymptotic Statistics* do van der Vaart[64] e no já mencionado *Econometrics* do Hayashi.[65] Os três primeiros desses resultados garantem que os modos de convergência estudados são "bem comportados" mediante transformações contínuas (em particular as operações algébricas usuais dos espaços Euclidianos), e estabelecem algumas relações entre esses modos de convergência.

[64] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Univ. Press, 1998

[65] Fumio Hayashi. *Econometrics*. Princeton University Press, 2000

Teorema 25 (Continuous mapping theorem)**.** Sejam $\boldsymbol{z}_n$, $n \geq 1$, e $\boldsymbol{z}$ vetores aleatórios em $\mathbb{R}^k$. Suponha que h é uma função de $\mathbb{R}^k$ em $\mathbb{R}^d$ cujo conjunto de pontos de continuidade $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ satisfaz $\mathbb{P}(\boldsymbol{z} \in \boldsymbol{\Theta}) = 1$. Então vale o seguinte:

1. Se $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{z}$, então $h \circ \boldsymbol{z}_n \xrightarrow{p} h \circ \boldsymbol{z}$.
2. Se $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$, então $h \circ \boldsymbol{z}_n \xrightarrow{d} h \circ \boldsymbol{z}$.

Nos teorema e corolário abaixo, estão implícitas as dimensões dos espaços correspondentes a cada um dos vetores aleatórios que ali aparecem. Essas dimensões podem variar de um item para outro de forma que as expressões façam sentido.

Teorema 26. Sejam $\boldsymbol{z}_n$, $\boldsymbol{w}_n$, $n \geq 1$, e $\boldsymbol{z}$ vetores aleatórios e seja $\boldsymbol{\xi}$ um vetor constante. Então vale o seguinte:

1. Se $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{z}$, então $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$.
2. Se $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{\xi}$, então $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{\xi}$.
3. Se $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ e $|\boldsymbol{z}_n - \boldsymbol{w}_n| \xrightarrow{p} 0$, então $\boldsymbol{w}_n \xrightarrow{d} \boldsymbol{z}$.
4. Se $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ e $\boldsymbol{w}_n \xrightarrow{p} \boldsymbol{\xi}$, então $(\boldsymbol{z}_n' \quad \boldsymbol{w}_n')' \xrightarrow{d} (\boldsymbol{z}' \quad \boldsymbol{\xi}')'$.

Corolário 27 (Lema de Slutsky)**.** Nas condições do Teorema 26, se $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ e $\boldsymbol{w}_n \xrightarrow{d} \boldsymbol{\xi}$, então

1. $\boldsymbol{z}_n + \boldsymbol{w}_n \xrightarrow{d} \boldsymbol{z} + \boldsymbol{\xi}$.
2. $\boldsymbol{w}_n' \boldsymbol{z}_n \xrightarrow{d} \boldsymbol{\xi}' \boldsymbol{z}$.
3. $\boldsymbol{z}_n / w_{n,1} \xrightarrow{d} \boldsymbol{z}/\xi_1$ desde que $\xi_1 \neq 0$.

O método Delta

A ideia do método Delta é utilizar uma expansão de Taylor para aproximar a distribuição de um vetor aleatório da forma $h \circ \boldsymbol{z}_n$. Nessa seção, para mantermo-nos fiéis à notação mais usual e a despeito da ressalva que fizemos na seção introdutória destas notas, escreveremos provisoriamente $h(\boldsymbol{z}_n)$ ao invés de $h \circ \boldsymbol{z}_n$. Seja, então, $h = (h_1 \quad h_2 \quad \cdots \quad h_d)'$ uma transformação com domínio $\Theta \subseteq \mathbb{R}^k$ e contradomínio $\mathbb{R}^d$. Suponha que as derivadas parciais

$$\partial_j^\ell h(\boldsymbol{\theta}) \coloneqq \lim_{\delta \to 0} \delta^{-1} \big(h_\ell(\boldsymbol{\theta} + \delta \boldsymbol{e}_j) - h_\ell(\boldsymbol{\theta}) \big)$$

existem e são contínuas numa vizinhança[66] $U \subseteq \Theta$ do ponto $\boldsymbol{\theta} \in \Theta$ para $j \in \{1, \ldots, k\}$ e $\ell \in \{1, \ldots, d\}$, onde $\boldsymbol{e}_j$ é o j–ésimo elemento da base canônica de $\mathbb{R}^k$ — por exemplo,

$$\boldsymbol{e}_2 = (0 \quad 1 \quad 0 \quad \cdots \quad 0)'$$

etc. Defina $\partial h_{\boldsymbol{\theta}}$ como sendo a matriz $(d \times k)$ cuja componente (j, ℓ) é $\partial_j^\ell h(\boldsymbol{\theta})$. Nessas condições, temos

$$h(\boldsymbol{\theta} + \boldsymbol{\xi}) = h(\boldsymbol{\theta}) + \partial h_{\boldsymbol{\theta}} \boldsymbol{\xi} + r(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \mathbb{R}^k, \tag{23}$$

em que $\lim_{\boldsymbol{\xi} \to \mathbf{0}} r(\boldsymbol{\xi}) = \mathbf{0}$.[67]

No enunciado abaixo, a sequência $(a_n)_{n \geq 1}$ é usualmente tomada com $a_n = \sqrt{n}$, mas não há razão para não sermos mais generalistas:

Teorema 28 (Método Delta)**.** Seja h como introduzida acima. Sejam $(\boldsymbol{z}_n)_{n \geq 1}$ uma sequência de vetores aleatórios em $\mathbb{R}^k$, com

[66] Usaremos o termo *vizinhança* de um ponto $\boldsymbol{\theta}$ no seguinte sentido: para $\delta > 0$ suficientemente pequeno, os conjuntos $\text{ball}(\boldsymbol{\theta}; \delta)$ estão inteiramente contidos em U.

[67] Veja o exercício 64.

$\mathbb{P}(\boldsymbol{z}_n \in \Theta) = 1$ para todo n, e $(a_n)_{n\geq 1}$ uma sequência de números reais com $a_n \to +\infty$. Se $a_n(\boldsymbol{z}_n - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{z}$, então

$$a_n\big(h(\boldsymbol{z}_n) - h(\boldsymbol{\theta})\big) \xrightarrow{d} \partial h_{\boldsymbol{\theta}} \boldsymbol{z}.$$

Ademais,

$$\Big| a_n\big(h(\boldsymbol{z}_n) - h(\boldsymbol{\theta})\big) - \partial h_{\boldsymbol{\theta}}\big(a_n(\boldsymbol{z}_n - \boldsymbol{\theta})\big) \Big| \xrightarrow{p} 0.$$

Os Reis do infinito

A *Lei Forte dos Grandes Números* e o *Teorema Central do Limite* são os dois principais resultados da Probabilidade Axiomática; reinam, absolutos, sobre toda a teoria assintótica. Abaixo enunciamos esses resultados, em sua versão multivariada, para sequências iid: para modelos em que assume-se um sistema de amostragem *cross section*, isso é tudo que precisaremos.

Teorema 29 (2ª Lei Forte dos Grandes Números de Kolmogorov)**.** Seja $(\boldsymbol{z}_n)_{n\geq 1}$ uma sequência iid de vetores aleatórios em $\mathbb{R}^k$, com $\mathbb{E}|\boldsymbol{z}_n| < \infty$. Então

$$\mathbb{P}\Big\{\omega \in \Omega\colon \lim_{n\to\infty} \big|\bar{\boldsymbol{z}}_n(\omega) - \mathbb{E}(\boldsymbol{z}_1)\big| = 0\Big\} = 1,$$

em que $\bar{\boldsymbol{z}}_n \coloneqq n^{-1}\sum_{i=1}^n \boldsymbol{z}_i$.

O vetor aleatório $\bar{\boldsymbol{z}}_n \coloneqq n^{-1}\sum_{i=1}^n \boldsymbol{z}_i$ que aparece no teorema acima é chamado a **média amostral** (da amostra $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$). O teorema afirma que, com $\mathbb{P}$–probabilidade 1, a média amostral converge para o valor esperado (comum) dos vetores aleatórios $\boldsymbol{z}_n$, $n \geq 1$. Nos exercícios abaixo o leitor será convidado a provar que, sob essas condições, vale também a convergência $\bar{\boldsymbol{z}}_n \xrightarrow{p} \mathbb{E}(\boldsymbol{z}_1)$.

Teorema 30 (Teorema Central do Limite de Lindeberg–Levy)**.** Seja $(\boldsymbol{z}_n)_{n\geq 1}$ uma sequência iid de vetores aleatórios em $\mathbb{R}^k$. Suponha que $\mathbb{E}(\boldsymbol{z}_1'\boldsymbol{z}_1) < \infty$. Então

$$\sqrt{n}\big(\bar{\boldsymbol{z}}_n - \boldsymbol{\mu}\big) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(\boldsymbol{z}_i - \boldsymbol{\mu}\big) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{24}$$

onde $\boldsymbol{\mu} \coloneqq \mathbb{E}(\boldsymbol{z}_1)$ e $\boldsymbol{\Sigma} \coloneqq \mathbb{V}(\boldsymbol{z}_1)$.

O teorema acima pode não parecer tão impressionante à primeira vista, mas sua força torna-se mais evidente após um segundo olhar. Primeiramente, note que podemos reinterpretar a convergência em (24) como uma afirmação de que a média amostral $\bar{\boldsymbol{z}}_n$ possui distribuição aproximadamente $\mathrm{N}(\boldsymbol{\mu}, n^{-1/2}\boldsymbol{\Sigma})$, desde que n seja "suficientemente grande". Pensemos no caso univariado (i.e., $k = 1$) para ilustrar: nesse caso, escrevendo z_n para a primeira componente do vetor $\boldsymbol{z}_n$, a condição $\mathbb{E}(\boldsymbol{z}_1'\boldsymbol{z}_1) < \infty$ se resume a $\mathbb{E}(z_1^2) < \infty$, e já sabemos que isso é equivalente à exigência de que $\sigma^2 \coloneqq \mathbb{V}(z_1) < \infty$. Escrevendo ainda $\mu \coloneqq \mathbb{E}(z_1)$, o Teorema Central do Limite nos dá a aproximação

$$\mathbb{P}(\bar{z}_n \leq \xi) \approx \int_{-\infty}^{\xi} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(t-\mu)^2/(2\sigma^2)}\,\mathrm{d}t, \quad \xi \in \mathbb{R},$$

para n suficientemente grande. Veja bem, as únicas hipóteses feitas sobre a distribuição conjunta das variáveis aleatórias $z_1, \ldots, z_n$ é que estas são independentes, identicamente distribuídas, e que possuem variâncias finitas. Afora isso, a distribuição marginal das variáveis aleatórias z_i é arbitrária. O surpreendente é que essa arbitrariedade não importa quando olhamos para a distribuição de probabilidades da média amostral $\bar{z}_n$: essa distribuição é aproximadamente normal. Dito de outra forma, ao tomar uma média aritmética de variáveis aleatórias quadrado–integráveis, independentes e identicamente distribuídas, ocorre que não importa qual seja a distribuição marginal subjacente, essa média necessariamente se distribuirá de acordo com uma lei muito próxima da distribuição normal!

Exercícios

Exercício 46. Sejam $\boldsymbol{z}_n$, $n \geq 1$ e $\boldsymbol{z}$ vetores aleatórios em $\mathbb{R}^k$, e seja $(a_n)_{n\geq 1}$ uma sequência de números reais positivos. Escrevemos $\boldsymbol{z}_n = \boldsymbol{z} + o_{\mathbb{P}}(a_n)$ se, e somente se, $(\boldsymbol{z}_n - \boldsymbol{z})/a_n \to 0$ em probabilidade.[68] Verifique que são equivalentes as seguintes afirmações:

1. $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{z}$.
2. $\boldsymbol{z}_n - \boldsymbol{z} \xrightarrow{p} \mathbf{0} \in \mathbb{R}^k$.
3. $\left|\boldsymbol{z}_n - \boldsymbol{z}\right| \xrightarrow{p} 0 \in \mathbb{R}$.
4. Para todo $\delta > 0$, $\mathbb{P}\big(\left|\boldsymbol{z}_n - \boldsymbol{z}\right| > \delta\big) \to 0 \in \mathbb{R}$.[69]

[68] Há aqui um ligeiro abuso de notação — seria mais preciso escrever, e.g., $\boldsymbol{z}_n - \boldsymbol{z} \in o_{\mathbb{P}}(a_n)$ etc — mas, em que pese essa incorreção, um que é extremamente útil.

[69] Lembre que uma sequência

$$(\boldsymbol{\xi}_n)_{n\geq 1} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots)$$

em $\mathbb{R}^k$ **converge** para um vetor $\boldsymbol{\xi} \in \mathbb{R}^k$ se, e somente se, para todo $\delta > 0$ existe um número natural n_δ tal que a desigualdade $|\boldsymbol{\xi}_n - \boldsymbol{\xi}| < \delta$ é válida para todo $n \geq n_\delta$.

5. Para todo par de números positivos δ e δ' existe um natural $n(\delta, \delta')$ tal que $\sup_{n \geq n(\delta,\delta')} \mathbb{P}\big(\big|\boldsymbol{z}_n - \boldsymbol{z}\big| > \delta\big) < \delta'$.
6. Para todo $\delta > 0$, $\mathbb{P}\big(\big|\boldsymbol{z}_n - \boldsymbol{z}\big| \leq \delta\big) \to 1 \in \mathbb{R}$.
7. $\boldsymbol{z}_n = \boldsymbol{z} + o_{\mathbb{P}}(1)$.

Exercício 47. Nas mesmas condições do exercício acima, escrevemos $\boldsymbol{z}_n = \boldsymbol{z} + O_{\mathbb{P}}(a_n)$ se, e somente se, para todo $\delta > 0$ existir uma constante $C_\delta > 0$ tal que

$$\mathbb{P}\left(\frac{|\boldsymbol{z}_n - \boldsymbol{z}|}{a_n} > C_\delta\right) < \delta$$

para todo $n \geq 1$. Mostre que

1. $o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$.
2. $o_{\mathbb{P}}(1) + O_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$.
3. $o_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$.
4. $\big(1 + o_{\mathbb{P}}(1)\big)^{-1} = O_{\mathbb{P}}(1)$.
5. $o_{\mathbb{P}}(a_n) = a_n \cdot o_{\mathbb{P}}(1)$.
6. $O_{\mathbb{P}}(a_n) = a_n \cdot O_{\mathbb{P}}(1)$.
7. $o_{\mathbb{P}}(O_{\mathbb{P}}(1)) = o_{\mathbb{P}}(1)$.

(Conforme mencionado acima, há um abuso de notação aqui, e as "igualdades" representam, na verdade, uma implicação — isto é, essas igualdades devem ser lidas "da esquerda para a direita". Preencha os detalhes relativos a notação.)

Exercício 48. *Prove que $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ se, e somente se, vale o seguinte: para toda função $h\colon \mathbb{R}^k \to \mathbb{R}$ contínua e limitada[70] vale $\lim_{n\to\infty} \mathbb{E}(h \circ \boldsymbol{z}_n) = \mathbb{E}(h \circ \boldsymbol{z})$.

[70] h ser limitada significa que $\sup_{\boldsymbol{\xi} \in \mathbb{R}^k} |h(\boldsymbol{\xi})| < \infty$ ou, o que é dizer o mesmo, que existe uma constante $c > 0$ tal que $|h(\boldsymbol{\xi})| \leq c$ para todo $\boldsymbol{\xi} \in \mathbb{R}^k$.

Exercício 49. Dizemos que $(\boldsymbol{z}_n)_{n\geq 1}$ **converge quase certamente** para $\boldsymbol{z}$, e que $\boldsymbol{z}$ é o **limite quase certo** da sequência $(\boldsymbol{z}_n)_{n\geq 1}$ se, e somente se, valer

$$\mathbb{P}\Big\{\omega \in \Omega\colon \lim_{n\to\infty} \boldsymbol{z}_n(\omega) = \boldsymbol{z}(\omega)\Big\} = 1.$$

Notação: $\boldsymbol{z}_n \xrightarrow{q.c.} \boldsymbol{z}$. Já vimos que a sequência de variáveis aleatórias $(z_n)_{n\geq 1}$ do exemplo 22 converge em probabilidade para 0 (dê os detalhes se não estiver convencido). Mostre que $z_n \not\to 0$ quase certamente.

Exercício 50. Seja $q \geq 1$. Dizemos que a sequência $(\boldsymbol{z}_n)_{n\geq 1}$ **converge em q–norma** para $\boldsymbol{z}$ se, e somente se, $\mathbb{E}\big(|\boldsymbol{z}_n - \boldsymbol{z}|^q\big) \to 0$. Notação: $\boldsymbol{z}_n \xrightarrow{q} \boldsymbol{z}$, "$\boldsymbol{z}_n \to \boldsymbol{z}$ em q–norma" etc.

1. Mostre que, se $\boldsymbol{z}_n \to \boldsymbol{z}$ em q–norma, então $\boldsymbol{z}_n \to \boldsymbol{z}$ em probabilidade.
2. Nas mesmas condições do exemplo 22, sejam w_n as variáveis aleatórias definidas por $w_n(\omega) \coloneqq 2^n \cdot z_n(\omega)$, $n \geq 1$. Verifique que $w_n \xrightarrow{p} 0$ mas $\mathbb{E}w_n \to +\infty$.[71] Conclua que convergência em probabilidade não implica convergência em q–norma.

[71] Se $(a_n)_{n\geq 1}$ é uma sequência de números reais, escrevemos $a_n \to +\infty$ se, e somente se, $(1+a_n)^{-1} \to 0$.

Exercício 51. Mostre que, se $\boldsymbol{z}_n \xrightarrow{q.c.} \boldsymbol{z}$, então $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{z}$.

Exercício 52. Mostre que, se $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{z}$ e $\boldsymbol{w}_n \xrightarrow{p} \boldsymbol{w}$, então

$$(\boldsymbol{z}_n' \quad \boldsymbol{w}_n')' \xrightarrow{p} (\boldsymbol{z}' \quad \boldsymbol{w}')'.$$

Exercício 53. Demonstre o Teorema 25.

Exercício 54. *Demonstre o Teorema 26.

Exercício 55. Os modos de convergência introduzidos para sequências de vetores aleatórios podem ser facilmente estendidos para sequências de matrizes aleatórias, utilizando a aplicação τ do exercício 16. Seja $(\boldsymbol{A}_n)_{n\geq 1}$ uma sequência de matrizes aleatórias, e seja $\boldsymbol{A}$ uma matriz *não aleatória*.[72] Mostre que:

[72] Preencha os detalhes quanto às dimensões das matrizes e vetores envolvidos. Essas dimensões podem depender de n?

1. Se $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ e $\boldsymbol{A}_n \xrightarrow{p} \boldsymbol{A}$, então $\boldsymbol{A}_n\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{A}\boldsymbol{z}$.
2. Nas condições do item acima, se adicionalmente $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, então $\boldsymbol{A}_n\boldsymbol{z}_n \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$.
3. Se $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ e $\boldsymbol{A}_n \xrightarrow{p} \boldsymbol{A}$, com $\boldsymbol{A}$ invertível, então $\boldsymbol{A}_n^{-1}\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{A}^{-1}\boldsymbol{z}$ e $\boldsymbol{z}_n'\boldsymbol{A}_n^{-1}\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}'\boldsymbol{A}^{-1}\boldsymbol{z}$.

Exercício 56. Prove a seguinte *lei fraca dos grandes números*: se $(\boldsymbol{z}_n)_{n\geq 1}$ é uma sequência iid de vetores aleatórios em $\mathbb{R}^k$ com $\mathbb{E}(\boldsymbol{z}_n'\boldsymbol{z}_n) < \infty$, então $\bar{\boldsymbol{z}}_n \xrightarrow{p} \mathbb{E}(\boldsymbol{z}_1)$, onde $\bar{\boldsymbol{z}}_n \coloneqq n^{-1}\sum_{i=1}^n \boldsymbol{z}_i$. Dica: comece com o caso $\mathbb{E}(\boldsymbol{z}_1) = \boldsymbol{0}$ e use a *desigualdade de Markov*: se z é uma variável aleatória não negativa, então $\delta\, \mathbb{P}(z \geq \delta) \leq \mathbb{E}(z)$.

Exercício 57. Com a mesma notação do exercício acima (mas sem a hipótese de que a sequência é iid), prove a seguinte *lei fraca dos grandes números*: se $\mathbb{E}(\bar{\boldsymbol{z}}_n) \to \boldsymbol{\mu} \in \mathbb{R}^k$ e $\mathbb{V}(\bar{\boldsymbol{z}}_n) \to \boldsymbol{0}$, então $\bar{\boldsymbol{z}}_n \xrightarrow{p} \boldsymbol{\mu}$.

Exercício 58. Suponha que $(z_n)_{n\geq 1}$ é uma sequência iid de variáveis aleatórias com distribuição normal padrão. Verifique que $z_n \xrightarrow{d} z_1$ mas $z_n \not\to z_1$ em probabilidade.

Exercício 59. Seja $(z_n)_{n\geq 1}$ uma sequência de variáveis aleatórias com

$$\mathbb{P}(z_n \leq \xi) = \int_{-\infty}^{\xi} \frac{\Gamma\big((n+1)/2\big)}{\sqrt{n\pi}\,\Gamma\big(n/2\big)} (1+t^2/n)^{-(n+1)/2} \mathrm{d}t, \quad \xi \in \mathbb{R},\, n \geq 1,$$

onde Γ é a *função gamma*,

$$\Gamma(\xi) = \int_0^{\infty} t^{\xi-1} \mathrm{e}^{-t}\, \mathrm{d}t, \quad \xi \in \mathbb{R}.$$

Isto é, para $n \geq 1$, a variável aleatória z_n tem distribuição Student–t com n graus de liberdade. Mostre que $z_n \xrightarrow{d} \mathrm{N}(0,1)$.

Exercício 60. Em que sentido é verdade que $o_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$? E quanto à igualdade $O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$, podemos afirmar que ela é verdadeira?

Exercício 61. Fixado $\boldsymbol{\theta} \in \mathbb{R}^k$, mostre que, se $\sqrt{n}(\boldsymbol{z}_n - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{z}$, então $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{\theta}$.

Exercício 62. Fixado $\boldsymbol{\theta} \in \mathbb{R}^k$, mostre que, se $\mathbb{E}(\boldsymbol{z}_n) \to \boldsymbol{\theta}$ e $|\mathbb{V}(\boldsymbol{z}_n)| \to 0$, então $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{\theta}$. Faça o caso $k = 1$ primeiro.

Exercício 63. Prove o seguinte: se, para todo $\delta > 0$ a série $\sum_{n=1}^{\infty} \mathbb{P}(|\boldsymbol{z}_n| > \delta)$ converge, então $\boldsymbol{z}_n \xrightarrow{q.c.} \mathbf{0}$.

Exercício 64. Mostre que o resto $r(\cdot)$ na equação (23) satisfaz $\lim_{\boldsymbol{\xi}\to\mathbf{0}} r(\boldsymbol{\xi}) = \mathbf{0}$.

Exercício 65. *Prove o Teorema 28.

Exercício 66. *Mostre que, se o Teorema Central do Limite de Lindeberg–Levy é válido para $k = 1$, então o teorema é válido para todo $k \geq 1$.

Consistência e Gaussianidade assintóticas

"A skeptic, I would ask for consistency first of all."

Sylvia Plath, *The Unabridged Journals of Sylvia Plath.*

Sylvia Plath.

Conta a anedota que o ganhador do prêmio Nobel de Economia Clive W. J. Granger certa vez teria dito: "*If you can't get it right as n goes to infinity you shouldn't be in this business.*" Felizmente nossos esforços até aqui não foram em vão, e veremos que — sob certas condições — o estimador de mínimos quadrados ordinários *gets it right* ao $n \to \infty$. Enunciaremos os teoremas de consistência e Normalidade assintótica com bastante generalidade, para que o leitor tenha na manga um coringa a ser usado em cenários menos restritivos do que aquele sobre o qual essas notas se concentram (a saber, a situação em que os dados são *cross section*). Nessa seção, consideraremos que está dada uma sequência (infinita) $\{(y_i, \boldsymbol{x}_i)\}_{i\geq 1}$ onde, como antes, os y_i's são variáveis aleatórias reais e os $\boldsymbol{x}_i$'s são vetores aleatórios em $\mathbb{R}^{d+1}$. Também como antes, $\boldsymbol{X}$ denota a matriz $n \times (d+1)$ cuja componente ij é x_{ij}, e $\boldsymbol{y}$ denota o vetor $n \times 1$ cuja i-ésima componente é y_i.

Consistência

Um requerimento que parece ser o mínimo a se fazer quando desejamos desenvolver a teoria assintótica para o estimador de MQO é que a sequência $\{(y_i, \boldsymbol{x}_i)\}_{i\geq 1}$ seja **estacionária**,[73] no seguinte sentido: para qualquer $n \geq 1$, qualquer $t \geq 1$ e quaisquer subconjuntos

[73] A noção que introduzimos aqui é a de *estacionariedade forte*, em contraposição à noção de estacionariedade fraca (ou *em covariância*) que é prevalente em textos de séries temporais. Observe que a definição depende da medida de probabilidade $\mathbb{P}$.

Borelianos $A_1, A_2, \ldots, A_n \subseteq \mathbb{R} \times \mathbb{R}^{d+1}$, vale a igualdade

$$\mathbb{P}\big((y_1, \boldsymbol{x}_1) \in A_1, \ldots, (y_n, \boldsymbol{x}_n) \in A_n\big) = \mathbb{P}\big((y_{1+t}, \boldsymbol{x}_{1+t}) \in A_1, \ldots, (y_{n+t}, \boldsymbol{x}_{n+t}) \in A_n\big).$$

Considerando, por conveniência, que as unidades amostrais estão indexadas no "tempo", a identidade acima pode ser interpretada como uma exigência (ou uma suposição) de que o procedimento de amostragem[74] independa, do ponto de vista distribucional, do instante de tempo em que é iniciado. Em particular, sob estacionariedade vemos que amostras de tamanho n possuem sempre uma mesma distribuição conjunta subjacente, que os vetores aleatórios $(y_i, \boldsymbol{x}_i)$, $i \geq 1$, possuem uma distribuição marginal comum etc.

[74] O qual pode ser puramente observacional, é claro.

Seria fortuito se, com estacionariedade apenas, pudéssemos obter alguma propriedade assintótica interessante para o estimador $\widehat{\boldsymbol{\beta}}_n$, mas coisas interessantes raramente se revelam tão facilmente: será preciso introduzir mais estrutura na medida de probabilidade $\mathbb{P}$ se quisermos chegar a algum lugar. O próximo atributo a exigir é aquilo que chamamos de *ergodicidade* — um termo que soa misterioso mas que é, de fato, o requerimento mínimo para que se possa fazer qualquer inferência quanto à distribuição subjacente a um conjunto de dados observados: o requerimento de que, em amostras "grandes", observemos aproximadamente $n \times \mathbb{P}\big((y_1, \boldsymbol{x}_1) \in A\big)$ pontos amostrais na região $A \subseteq \mathbb{R} \times \mathbb{R}^{d+1}$. Dizemos que uma sequência estacionária $\{\boldsymbol{z}_i\}_{i\geq 1} \equiv \{(y_i, \boldsymbol{x}_i)\}_{i\geq 1}$ é **ergódica** se, e somente se, para qualquer $t \geq 1$ e quaisquer subconjuntos Borelianos $A_0, A_1, \ldots A_t \subseteq \mathbb{R} \times \mathbb{R}^{d+1}$, vale

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big[\boldsymbol{z}_i \in A_0, \ldots, \boldsymbol{z}_{i+t} \in A_t\big] \overset{q.c.}{\to} \mathbb{P}(\boldsymbol{z}_1 \in A_0, \ldots, \boldsymbol{z}_{1+t} \in A_t).$$

Uma fato importante e de fácil verificação é que sequências iid são ergódicas — ver as páginas 488 e 489 em Karlin & Taylor.[75] Portanto, os resultados que enunciamos abaixo têm validade também no caso de dados *cross section*.

[75] Samuel Karlin and Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press, Boston, 2nd edition, 1975

Com ergodicidade obtemos a consistência do estimador $\widehat{\boldsymbol{\beta}}_n$, mas antes de prosseguirmos convém recordar o Teorema 3, segundo o qual $\mathbb{E}\big(\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\big) = \mathbf{0}$ desde que $\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})$ seja invertível (e onde $\boldsymbol{\beta} \coloneqq \big(\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})\big)^{-1}\mathbb{E}(\boldsymbol{X}'\boldsymbol{y})$). Como, sob estacionariedade, valem[76] as

[76] Veja o exercício 68.

igualdades $\mathbb{E}(\boldsymbol{X}'\boldsymbol{X}) = n\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')$, $\mathbb{E}(\boldsymbol{X}'\boldsymbol{y}) = n\mathbb{E}(\boldsymbol{x}_1 y_1)$ e também

$$\boldsymbol{\beta} = \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1)\big)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1), \tag{25}$$

vemos, do mencionado teorema, que sob essas condições é satisfeita a identidade

$$\mathbb{E}\big(\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\beta)\big) = \mathbf{0}, \quad i \geq 1,$$

algo que no jargão é chamado de "regressores predeterminados". Isso é importante porque há uma linha alternativa através da qual se pode conduzir o argumento até chegarmos à consistência de $\widehat{\boldsymbol{\beta}}_n$: poderíamos equivalentemente colocar como uma *suposição* que valem as igualdades $y_i = \boldsymbol{x}_i'\boldsymbol{b} + v_i$, $i \geq 1$, para algum *vetor de parâmetros desconhecidos* $\boldsymbol{b}$ e alguma sequência $(v_i)_{i\geq 1}$ de *termos de erro não observáveis*, e após isso exigir que os regressores sejam predeterminados com respeito a essa sequência, isto é, que $\mathbb{E}(\boldsymbol{x}_i v_i) = \mathbf{0}$ para todo $i \geq 1$. Ocorre que, se a sequência $\{(y_i, \boldsymbol{x}_i)\}_{i\geq 1}$ é estacionária e se $\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1)$ é invertível, então essa última exigência somente pode ocorrer quando $\boldsymbol{b} = \boldsymbol{\beta}$, com $\boldsymbol{\beta}$ dado pela equação (25).

Estamos prontos para enunciar o teorema de consistência, mas antes de prosseguir, lembre que

$$\widehat{\boldsymbol{\beta}}_n = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \Big(\tfrac{1}{n}\sum\nolimits_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\Big)^{-1}\tfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i.$$

Teorema 31 (Consistência do estimador de MQO)**.** Suponha que $\{(y_i, \boldsymbol{x}_i)_{i\geq 1}\}$ é uma sequência ergódica, e que $\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')$ é invertível. Então

$$\widehat{\boldsymbol{\beta}}_n \overset{q.c.}{\to} \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1).$$

Sketch da demonstração. Uma adaptação do exercício 70 nos dá

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' \overset{p}{\to} \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1') \tag{26}$$

e, como o conjunto das matrizes invertíveis é um subconjunto aberto (verifique) do espaço $M\big((d+1)\times(d+1)\big)$, vemos que $\boldsymbol{X}'\boldsymbol{X}$ é invertível para n suficientemente grande. Pelo mesmo exercício temos a convergência $n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i y_i \overset{p}{\to} \mathbb{E}(\boldsymbol{x}_1 y_1)$, e isso é quase tudo que precisamos para obter $\widehat{\boldsymbol{\beta}} \overset{p}{\to} \boldsymbol{\beta}$: agora resta apenas aplicar o Teorema 25 algumas vezes. Para convergência quase certa, ver o Teorema 5.6, seção 9, em Karlin & Taylor.[77] ■

[77] Samuel Karlin and Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press, Boston, 2nd edition, 1975

Normalidade assintótica

Quanto à distribuição assintótica do estimador de mínimos quadrados ordinários, temos o seguinte:

Teorema 32 (Normalidade assintótica do estimador de MQO)**.** Nas condições do Teorema 31, seja $u_i \coloneqq y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, $i \geq 1$, onde $\boldsymbol{\beta} \coloneqq \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1)$. Assuma adicionalmente que

A1. $\mathbb{E}\big(|u_1^2\boldsymbol{x}_1\boldsymbol{x}_1'|\big) < \infty$.
A2. $\mathbb{E}\big(\boldsymbol{x}_{\ell+1}u_{\ell+1} \,\big|\, (\boldsymbol{x}_i, u_i), 1 \leq i \leq \ell\big) = \mathbf{0}$ para todo $\ell \geq 1$.

Então

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

onde $\boldsymbol{\Sigma} \coloneqq \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(u_1^2\boldsymbol{x}_1\boldsymbol{x}_1')\big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}$.

Demonstração. Ver a Proposição 2.1 no *Econometrics* do Hayashi[78] para a versão geral enunciada acima. Para o cenário iid, primeiramente observe que nesse caso (sob as suposições do Teorema 31) a condição *A2* é sempre satisfeita, de forma que somente precisamos impor *A1*. Note também que $n^{-1/2}\sum_{i=1}^n \boldsymbol{x}_i u_i \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, \mathbb{E}(u_1^2\boldsymbol{x}_1\boldsymbol{x}_1')\big)$, pelo Teorema 30. Ademais, no presente caso podemos obter a convergência em (26) usando a Lei Forte dos Grandes Números para sequências iid. Agora basta notar que

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) = \Big(\tfrac{1}{n}\sum\nolimits_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\Big)^{-1} \tfrac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i u_i.$$

e aplicar os teoremas vistos no capítulo anterior para obter a convergência desejada. ■

[78] Fumio Hayashi. *Econometrics*. Princeton University Press, 2000

Exercícios

Quaisquer erros nos enunciados são — por questões de consistência — intencionais.

Exercício 67. Mostre que, se a sequência $\{(y_i, \boldsymbol{x}_i)\}_{i\geq 1}$ é iid, então ela é estacionária.

Exercício 68. Mostre que, se a sequência $\{(y_i, \boldsymbol{x}_i)\}_{i\geq 1}$ é estacionária, então

1. $\mathbb{E}(\boldsymbol{X}'\boldsymbol{X}) = n\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')$.

2. $\mathbb{E}(\boldsymbol{X}'\boldsymbol{y}) = n\mathbb{E}(\boldsymbol{x}_1 y_1)$.

Conclua que, sob estacionariedade, vale $\boldsymbol{\beta} = \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1)$ e $\mathbb{E}(\boldsymbol{X}'\boldsymbol{u}) = n\mathbb{E}(\boldsymbol{x}_1 u_1) = \boldsymbol{0}$, onde $\boldsymbol{u}' = (u_1 \quad \cdots \quad u_n)$ com $u_i := y_i - \boldsymbol{x}_i'\beta$, $i \geq 1$.

Exercício 69. Suponha que a sequência $\{(y_i, \boldsymbol{x}_i)\}_{i\geq 1}$ é ergódica. Seja A um subconjunto Boreliano de $\mathbb{R} \times \mathbb{R}^{d+1}$, e ponha $\alpha := \mathbb{P}\big((y_1, \boldsymbol{x}_1) \in A\big)$. Mostre que, dado $\delta > 0$ arbitrário, podemos encontrar um evento Ω_δ com $\mathbb{P}(\Omega_\delta) = 1$ e um número natural n_δ tal que, para $n \geq n_\delta$ e $\omega \in \Omega_\delta$, a proporção dos pontos

$$\big(y_1(\omega), \boldsymbol{x}_1(\omega)\big), \ldots, \big(y_n(\omega), \boldsymbol{x}_n(\omega)\big)$$

que encontramos na região A é algo entre $\alpha - \delta$ e $\alpha + \delta$. Isto é, sob ergodicidade, o número de pontos em um *scaterplot* que se situam numa região A qualquer é aproximadamente $n \cdot \mathbb{P}\big((y_1, \boldsymbol{x}_1) \in A\big)$, desde que n seja suficientemente grande.

Exercício 70. Seja $h\colon \mathbb{R} \times \mathbb{R}^{d+1} \to \mathbb{R}$ uma função que pode ser expressa como uma combinação linear de funções indicadoras. Isto é, existem um $k \geq 1$, subconjuntos Borelianos $A_1, \ldots, A_k \subseteq \mathbb{R} \times \mathbb{R}^{d+1}$, e constantes $a_1, \ldots, a_k \in \mathbb{R}$ tais que

$$h(\xi, \boldsymbol{\zeta}) = \sum_{\ell=1}^{k} a_\ell \cdot \mathbb{I}\big[(\xi, \boldsymbol{\zeta}) \in A_\ell\big], \qquad (\xi, \boldsymbol{\zeta}) \in \mathbb{R} \times \mathbb{R}^{d+1}.$$

Mostre que, se a sequência $\{\boldsymbol{z}_i\}_{i\geq 1} \equiv \{(y_i, \boldsymbol{x}_i)\}_{i\geq 1}$ é ergódica, então

$$\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{z}_i) \overset{q.c.}{\to} \mathbb{E}h(\boldsymbol{z}_1).$$

Com algum esforço adicional, verifique que $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \overset{p}{\to} \mathbb{E}\boldsymbol{z}_i$, desde que $\mathbb{E}|\boldsymbol{z}_1| < \infty$.

Exercício 71. Dê os detalhes da demonstração do Teorema 31 (apenas para o caso de convergência em probabilidade).

Exercício 72. Dê os detalhes da demonstração do Teorema 32 (apenas para o caso iid).

Exercício 73. Mostre que a condição

A2'. $\mathbb{E}\big(\boldsymbol{x}_{\ell+1}u_{\ell+1} \,\big|\, \boldsymbol{x}_{\ell+1}, (\boldsymbol{x}_i, u_i), 1 \leq i \leq \ell\big) = \mathbf{0}$ para todo $\ell \geq 1$,

implica a condição *A2*.

Exercício 74. Nas condições do Teorema 32, se $\mathbb{P}(x_{1,0} = 1) = 1$ então $\mathbb{E}(u_{\ell+1} \,|\, u_\ell, \ldots, u_1) = 0$ para $\ell \geq 1$. Conclua que, nessas condições, os termos de erro $(u_i)_{i\geq 1}$ não apresentam autocorrelação serial.

Exercício 75. Verifique que, se $\mathbb{E}\big(|\boldsymbol{x}_1|^4\big)$ e $\mathbb{E}\big(|u_1|^4\big)$ são ambas finitas, então vale *A1* no Teorema 32.

Exercício 76. Compare a variância assintótica de $\widehat{\boldsymbol{\beta}}_n$ dada no Teorema 32 com a variância em amostras finitas que havíamos obtido no Teorema 19.

Sob a tese...

"El diccionario se basa en la hipótesis, obviamente no comprobada, de que los idiomas están compuestos de sinónimos equivalentes."

J. L. Borges.

J. L. Borges, 1969.

Introdução

Uma *hipótese estatística* nada mais é do que uma sentença (formal) envolvendo parâmetros de um modelo. Por exemplo, definindo $\boldsymbol{\beta} := \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1)$, fixada uma aplicação $h\colon \mathbb{R}^{d+1} \to \mathbb{R}^k$, e sendo dado $\boldsymbol{\theta} \in \mathbb{R}^k$ arbitrário, a sentença

$$\text{“}h(\boldsymbol{\beta}) = \boldsymbol{\theta}\text{”} \qquad (\mathrm{H}_0)$$

é uma hipótese estatística, usualmente chamada de *hipótese nula*. Note que, embora não explicitamente, a sentença H_0 acima refere-se à medida de probabilidade $\mathbb{P}$.[79] Por essa razão, vamos adotar um ligeiro abuso de notação aqui e identificar cada sentença com o conjunto que ela define, isto é, escreveremos, e.g.

$$\mathrm{H}_0 \equiv \{\mathbb{P} \in \mathrm{H}_* \colon \mathrm{H}_0 \text{ é satisfeita por } \mathbb{P}\}$$

etc, onde H_* denota uma hipótese/suposição que está implícita na discussão[80] (a qual assumimos como verdadeira e que não será testada). Por exemplo, H_* pode ser a suposição de que a amostra é iid, de que as variáveis aleatórias têm variâncias finitas etc.

Um *teste de hipóteses* é uma regra de decisão cujo espaço de escolha, $\mathscr{E}$, é binário; uma dessas escolhas é interpretada como *rejeição da hipótese* H_0 e a outra como *não-rejeição da hipótese* H_0.

[79] Pois $\boldsymbol{\beta}$ está definido em termos do operador $\mathbb{E}$, o qual por sua vez é definido em termos de $\mathbb{P}$.

[80] H_* exerce aqui o papel de *domínio de discurso* e garante que o conjunto H_0 especificado pelo *axioma de compreensão* está bem definido.

A regra de decisão é representada por uma função $D\colon \Omega \to \mathscr{E}$, em geral expressa na forma[81]

$$D(\omega) = \mathbb{I}(\boldsymbol{z}_n(\omega) \in \mathrm{R}_\alpha), \quad \omega \in \Omega,$$

onde $\boldsymbol{z}_n$ é uma *estatística de teste*[82] e onde, para um *nível de significância* $0 < \alpha < 1$, a *região de rejeição* $\mathrm{R}_\alpha \subseteq \mathbb{R}^k$ é escolhida de tal forma que valha a implicação

$$\mathbb{P} \in \mathrm{H}_* \wedge \mathrm{H}_0 \Rightarrow \mathbb{P}(\boldsymbol{z}_n \in \mathrm{R}_\alpha) \approx \alpha. \tag{27}$$

A pergunta que se apresenta naturalmente aqui é: "de onde vem a região R_α?" Ora, se pudermos obter, assumindo $\mathrm{H}_* \wedge \mathrm{H}_0$, a distribuição do vetor aleatório $\boldsymbol{z}_n$, então é claro que podemos encontrar uma região R_α como acima: basta tomar $\xi > 0$ suficientemente grande e definir R_α implicitamente através da igualdade de eventos

$$(\boldsymbol{z}_n \in \mathrm{R}_\alpha) = (|\boldsymbol{z}_n| > \xi).$$

De fato, se pudermos (como dito) obter a distribuição de $\boldsymbol{z}_n$ sob $\mathrm{H}_* \wedge \mathrm{H}_0$, então é evidente que a probabilidade do evento acima converge para zero ao $\xi \to \infty$ e, em particular, se a mencionada distribuição for contínua, é possível até mesmo encontrar um valor ξ_α para o qual a referida probabilidade é *igual* a α.

MAS SOMOS HUMANOS e, como tais, através de exemplos é que pensamos. Vejamos um, para ilustrar: em um contexto de regressão linear, consideremos a hipótese[83]

$$\mathrm{H}_0 = \{\mathbb{P} \in \mathrm{H}_*\colon \beta_1 = 0\} \tag{28}$$

O Teorema 32 nos garante que $\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, desde que as suposições (H_*) no seu enunciado estejam satisfeitas. Em particular, se esse for o caso *e também valer* H_0, então a variável aleatória

$$z_n \coloneqq \frac{\sqrt{n}\widehat{\beta}_{n1}}{\sqrt{\sigma_{11}}}, \tag{29}$$

onde $\sigma_{\ell j}$ denota a componente (ℓ, j) da matriz $\boldsymbol{\Sigma}$, tem distribuição limite aproximadamente Normal padrão. Daqui, utilizando a aproximação Normal, podemos (por exemplo) fixar $0 < \alpha < 1$ e definir

[81] Sem perda de generalidade, podemos identificar $\mathscr{E} \equiv \{0, 1\}$.

[82] Em geral, estatísticas de teste são variáveis (ou vetores) aleatórias que dependem de ω somente através da amostra $\big(\boldsymbol{y}(\omega), \boldsymbol{X}(\omega)\big)$. Por isso, a regra de decisão D também é, em geral, da forma $D = \phi(\boldsymbol{y}, \boldsymbol{X})$ para alguma função ϕ.

[83] Nesse contexto, onde $y_1 = \sum_{j=0}^{d} \beta_j x_{1j} + u_1$, tal hipótese é comumente interpretada como a afirmação de que "$x_{1,1}$ não exerce efeito linear sobre y_1."

nossa regra de decisão via

$$D(\omega) = \begin{cases} \text{rejeitar}, & \text{se } z_n(\omega) < -\xi_\alpha \text{ ou } z_n(\omega) > \xi_\alpha; \\ \text{não rejeitar}, & \text{caso contrário}, \end{cases}$$

onde $\xi_\alpha > 0$ é a única solução para a equação $F(\xi_\alpha) = 1-\alpha/2$ e onde F denota aqui a função de distribuição acumulada de uma variável aleatória Normal padrão. Quer dizer, nesse exemplo nossa região de rejeição é dada pela união de dois intervalos: $\mathrm{R}_\alpha = (-\infty, -\xi_\alpha) \cup (\xi_\alpha, +\infty)$. Se o leitor ainda anseia por concretude, podemos tomar $\alpha = 0.01$, como é costumeiro, e nesse caso temos $\xi_\alpha \approx 2.575829$, o que nos leva a rejeitar a hipótese nula se (e somente se), $|z_n^{\text{obs}}| > 2.575829$.

O teste descrito acima baseia-se em uma aproximação, e portanto é chamado de um *teste assintótico*.[84] Por essa razão, o nível α nesse contexto é chamado de *nível de significância nominal* do teste (em contraposição a um nível exato que seria obtenível se conhecêssemos a distribuição *exata* de z_n). Para justificar com mais detalhes a aproximação empregada acima, recorremos à Proposição 23, segundo a qual podemos tomar $\delta > 0$ tão pequeno quanto queiramos e tal que, para todo n suficientemente grande ($n \geq n_\delta$), valham as quotas

$$F(\xi) - \delta \;\leq\; \mathbb{P}(z_n \leq \xi) \;\leq\; F(\xi) + \delta, \qquad \xi \in \mathbb{R}.$$

Com algumas linhas de conta e usando a simetria da distribuição Normal, as duas desigualdades acima nos dão, para $\xi \geq 0$,

$$2\big(1 - F(\xi) - 2\delta\big) \leq \mathbb{P}(|z_n| > \xi) \leq 2\big(1 - F(\xi) + 2\delta\big). \tag{30}$$

Assim, tomando ξ_α como anteriormente, a equação (30) nos dá

$$\mathbb{P}(|z_n| > \xi_\alpha) = \alpha + \mathrm{r}(\delta),$$

com $|\mathrm{r}(\delta)| \leq 4\delta$, para qualquer $\alpha \in (0,1)$ e todo n suficientemente grande.[85] Segue que, para $\mathbb{P} \in \mathrm{H}_* \wedge \mathrm{H}_0$ e $n \geq n_\delta$, temos

$$\mathbb{P}(\text{rejeitar } \mathrm{H}_0) \approx \alpha.$$

O teste que acabamos de descrever se resume à regra de decisão "rejeitar se, e somente se, $|z_n^{\text{obs}}| > \xi_\alpha$". Com efeito, um teste de hipóteses pode ser interpretado como um algoritmo, constituído das seguintes etapas:

[84] É importante ressaltar também que esse é um exemplo de teste *bilateral*.

[85] A questão sobre "quão grande deve ser o tamanho da amostra para que valham as referidas aproximações?" é interessantíssima, mas foge da nossa linha de exposição. O leitor interessado pode pesquisar sobre o Teorema de Berry–Esseen e também sobre expansões de Edgeworth. É importante também perceber que o n_δ depende, implicitamente, da medida de probabilidade $\mathbb{P}$.

1. Formule uma hipótese H_0, declare um nível de significância $\alpha \in (0,1)$, determine a região de rejeição R_α correspondente e escolha uma estatística de teste $\boldsymbol{z}_n \colon \Omega \to \mathbb{R}^k$.
2. Colete dados (isto é, observe o resultado ω do experimento ou estudo observacional[86]) e calcule o valor correspondente da estatística de teste $\boldsymbol{z}_n^{\text{obs}} \equiv \boldsymbol{z}_n(\omega)$.
3. Se $\boldsymbol{z}_n^{\text{obs}} \in \mathrm{R}_\alpha$, rejeite a hipótese H_0; caso contrário, não rejeite.

[86] Ou, se a partir do experimento/procedimento observacional não for possível determinar o valor ω, observe por exemplo $(\boldsymbol{y}(\omega), \boldsymbol{X}(\omega))$ etc.

Esse procedimento não requer, a princípio, qualquer teoria que o justifique.[87] Todavia, a heurística segundo a qual propõe-se a operacionalização acima é a seguinte:

[87] De fato, do ponto de vista algorítmico, não há qualquer relação *a priori* entre a hipótese e a região de rejeição. Tampouco há qualquer menção a probabilidades.

1. Num primeiro momento, escolhemos a região de rejeição R_α de tal forma que a implicação (27) valha. Na prática, a região de rejeição tipicamente delimita uma magnitude a partir da qual a estatística de teste é considerada "grande demais", ou seja, temos $\boldsymbol{z}_n^{\text{obs}} \in \mathrm{R}_\alpha$ se, e somente se, $|\boldsymbol{z}_n^{\text{obs}}| > \xi_\alpha$ para algum $\xi_\alpha > 0$ o qual é escolhido de tal forma que a sentença formal

$$\mathbb{P} \in \mathrm{H}_* \wedge \mathrm{H}_0 \Rightarrow \mathbb{P}(|\boldsymbol{z}_n| > \xi_\alpha) \approx \alpha$$

   seja verdadeira.

2. Se observarmos $\boldsymbol{z}_n^{\text{obs}} \in \mathrm{R}_\alpha$ e se a hipótese H_0 for verdadeira, então essa observação pode ser considerada *atípica* (ou *extrema*) no sentido de que, sob $\mathrm{H}_* \wedge \mathrm{H}_0$, é uma observação situada numa região que está na "cauda" da distribuição de probabilidades da estatística de teste, conforme a implicação no item anterior (lembre que α é tipicamente um número "pequeno"). Por essa razão — e supondo que não queiramos (ou que não tenhamos motivos para) abrir mão de H_* — consideramos nessas condições que a hipótese H_0 é "inverossímil" e que é "razoável" rejeitá-la.[88]

[88] A sentença $\neg\mathrm{H}_0$ é usualmente chamada de *hipótese alternativa* e denotada por H_1. No jargão, dizemos que "rejeitamos H_0 *em favor* de H_1".

3. Se, por outro lado, observamos $\boldsymbol{z}_n^{\text{obs}} \notin \mathrm{R}_\alpha$, então consideramos que "não há evidências suficientes para rejeitar a hipótese nula", no sentido de que essa observação é "compatível" com H_0:

$$\mathbb{P} \in \mathrm{H}_* \wedge \mathrm{H}_0 \Rightarrow \mathbb{P}(\boldsymbol{z}_n \notin \mathrm{R}_\alpha) \approx 1 - \alpha.$$

Com isso encerramos essa breve (e incompleta) recapitulação da teoria de testes de hipóteses. O leitor pode encontrar uma exposição

suficientemente rigorosa e acessível no livro clássico de Casella & Berger.[89] Agora, vamos passar ao problema de encontrar um estimador consistente para $\boldsymbol{\Sigma}$.

[89] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002

Um estimador para a covariância assintótica

Nosso objetivo nessa seção é propor um estimador $\widehat{\boldsymbol{\Sigma}}_n$ satisfazendo $\widehat{\boldsymbol{\Sigma}}_n \xrightarrow{p} \boldsymbol{\Sigma}$, isto é, queremos encontrar um estimador consistente para a matriz de covariância assintótica[90] do estimador de mínimos quadrados ordinários, a qual é dada por

$$\boldsymbol{\Sigma} \coloneqq \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(u_1^2\boldsymbol{x}_1\boldsymbol{x}_1')\big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}. \tag{31}$$

[90] O leitor mais atento certamente deu-se conta de que a estatística de teste (29) não é computável, a menos que o desvio padrão σ_{11} seja conhecido.

A motivação é a seguinte: o Teorema 32 nos dá, sob certas condições, a convergência $\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Em particular, se $\boldsymbol{\Sigma}$ é positiva definida, então vale que

$$\sqrt{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{Id}), \tag{32}$$

onde $\boldsymbol{\Sigma}^{\frac{1}{2}}$ é a única matriz positiva definida $\boldsymbol{A}$ que satisfaz $\boldsymbol{A}^2 = \boldsymbol{\Sigma}$. Nessas condições, se pudermos encontrar um estimador $\widehat{\boldsymbol{\Sigma}}_n$ como acima, então uma simples aplicação de um dos resultados derivados no exercício 55 nos garante que podemos substituir $\boldsymbol{\Sigma}$ por $\widehat{\boldsymbol{\Sigma}}_n$ em (32), e com essa substituição obtemos uma estatística de teste cujo valor é computável sempre que estivermos diante de uma hipótese nula da forma $\mathrm{H}_0 = \{\mathbb{P}\colon \boldsymbol{\beta} = \boldsymbol{b}\}$.

Agora, uma rápida inspeção da expressão (31) já sugere (ou, a essa altura, já deveria sugerir) como estimar $\boldsymbol{\Sigma}$: temos dois "ingredientes", o primeiro deles sendo $n^{-1}\boldsymbol{X}'\boldsymbol{X}$, o qual satisfaz

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' \xrightarrow{p} \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')$$

se $(\boldsymbol{x}_i)_{i\geq 1}$ é ergódica (com respeito a $\mathbb{P}$).[91] O segundo desses ingredientes seria, idealmente, a média amostral $n^{-1}\sum_{i=1}^{n} u_i^2\boldsymbol{x}_i\boldsymbol{x}_i'$, pois sob ergodicidade temos

$$\frac{1}{n}\sum_{i=1}^{n} u_i^2\boldsymbol{x}_i\boldsymbol{x}_i' \xrightarrow{p} \mathbb{E}(u_1^2\boldsymbol{x}_1\boldsymbol{x}_1'),$$

desde que $\mathbb{E}\big|u_1^2\boldsymbol{x}_1\boldsymbol{x}_1'\big| < \infty$. O problema aqui é que os termos de erro u_i não são observáveis e, portanto, o estimador que aparece

[91] E, note, vale também que $\big(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\big)^{-1}$ converge em probabilidade para $\big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}$.

no lado esquerdo na expressão acima não é computável a partir da amostra $(\boldsymbol{y}, \boldsymbol{X})$. Bom, na verdade isso não é de fato um problema, pois temos o seguinte:

Teorema 33. Suponha que $\{(y_i, \boldsymbol{x}_i)_{i\geq 1}\}$ é uma sequência ergódica satisfazendo as seguintes condições:

1. $\mathbb{E}\big(|\boldsymbol{x}_1\boldsymbol{x}_1'|^2\big) < \infty$.
2. $\mathbb{E}\big(|u_1^2\boldsymbol{x}_1\boldsymbol{x}_1'|\big) < \infty$.

Então[92] $n^{-1}\sum_{i=1}^n \widehat{u}_i^2\boldsymbol{x}_i\boldsymbol{x}_i' \xrightarrow{p} \mathbb{E}(u_1^2\boldsymbol{x}_1\boldsymbol{x}_1')$.

[92] Nesse enunciado, adotamos a mesma notação já usada anteriormente: definimos $\widehat{u}_i \coloneqq y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ e $u_i \coloneqq y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, onde $\widehat{\boldsymbol{\beta}}$ é o estimador de mínimos quadrados ordinários e onde $\boldsymbol{\beta} = [\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')]^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1)$.

Corolário 34. Nas condições do teorema acima, suponha adicionalmente que $\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')$ é invertível. Então

$$\widehat{\boldsymbol{\Sigma}}_n \xrightarrow{p} \boldsymbol{\Sigma},$$

onde $\widehat{\boldsymbol{\Sigma}}_n \coloneqq n(\boldsymbol{X}'\boldsymbol{X})^{-1}\big(\sum_{i=1}^n \widehat{u}_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\big)(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Exemplo 35. Vamos explorar novamente, agora com um pouco mais de generalidade, o exemplo visto na seção introdutória deste capítulo. Seja $\widehat{\sigma}_{\ell j}^2$ a componente (ℓ, j) da matriz $\widehat{\boldsymbol{\Sigma}}_n$ introduzida acima.[93] Consideremos a hipótese nula $\mathrm{H}_0 = \{\mathbb{P}\colon \beta_k = b\}$, onde estão fixados $0 \leq k \leq d$ e $b \in \mathbb{R}$. Satisfeitas as devidas hipóteses (H_*) temos, do Teorema 32, que

$$\mathbb{P} \in \mathrm{H}_* \wedge \mathrm{H}_0 \Rightarrow \frac{\sqrt{n}\big(\widehat{\beta}_{nk} - b\big)}{\widehat{\sigma}_{kk}} \xrightarrow{d} \mathrm{N}(0,1).$$

A importância disso reside, como dissemos, no fato de que a estatística de teste acima é da forma $z_n(\omega) = h\big(\boldsymbol{y}(\omega), \boldsymbol{X}(\omega), b\big)$, onde h é uma função "universal", no sentido de que independe de quaisquer parâmetros associados a $\mathbb{P}$. Isso quer dizer que *podemos computar o valor* $z_n(\omega)$ *para qualquer amostra* $\big(\boldsymbol{y}(\omega), \boldsymbol{X}(\omega)\big)$ *e qualquer valor hipotético* $b \in \mathbb{R}$.[94]

O exemplo acima é útil pois, como já mencionamos anteriormente, pode ser do interesse do pesquisador testar a hipótese de que um particular regressor, digamos x_{1k}, "não tem efeito linear sobre a resposta" $(\mathrm{H}_0 = \{\mathbb{P}\colon \beta_k = 0\})$. Todavia, uma situação mais realista é aquela em que gostaríamos de testar hipóteses sobre mais de um dos coeficientes em $\boldsymbol{\beta}$ (possivelmente todos); aí é preciso ter em mente que — conforme se aprende após alguma prática no

[93] Omitiremos o índice n em $\widehat{\sigma}_{\ell j}$ para evitar que a notação fique excessivamente carregada — o leitor pode considerar que, sempre que o símbolo "chapéu" der as caras na notação, implicitamente há um n ali escondido.

[94] Aqui parece ser um bom lugar para introduzir mais um jargão: a variável aleatória

$$\widehat{\mathrm{se}}(\widehat{\beta}_k) \coloneqq \frac{\widehat{\sigma}_{kk}}{\sqrt{n}}$$

é dita o **erro padrão de White**, ou o **erro padrão robusto a heteroscedasticidade**. Trata-se de um estimador para o desvio padrão $\sqrt{\mathbb{V}(\widehat{\beta}_k)}$.

"jogo" de teste de hipóteses — *não se deve testar múltiplas hipóteses isoladamente*. O teorema abaixo nos diz como proceder:

Teorema 36. Suponha que $\{(y_i, \boldsymbol{x}_i)_{i\geq 1}\}$ é uma sequência ergódica satisfazendo as seguintes condições ($\mathrm{H}_* =$ ergodicidade$\wedge$a$\wedge$b$\wedge$c$\wedge$d):

(a) $\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')$ é invertível.
(b) $\mathbb{E}\big(|u_1^2\boldsymbol{x}_1\boldsymbol{x}_1'|\big) < \infty$.
(c) $\mathbb{E}\big(\boldsymbol{x}_{\ell+1}u_{\ell+1} \,\big|\, (\boldsymbol{x}_i, u_i), 1 \leq i \leq \ell\big) = \mathbf{0}$ para todo $\ell \geq 1$.
(d) $\mathbb{E}(u_1^2\boldsymbol{x}_1\boldsymbol{x}_1')$ é invertível,

onde $u_i \coloneqq y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, com $\boldsymbol{\beta} = \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1)$. Sejam ainda $\boldsymbol{a}$ um vetor de $\mathbb{R}^k$ arbitrário mas fixado e $\boldsymbol{A}$ uma matriz não aleatória de dimensões $k \times (d+1)$, com $\operatorname{rank}(\boldsymbol{A}) = k$, e considere a hipótese nula $\mathrm{H}_0 = \{\mathbb{P}\colon\ \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{a}\}$. Então, para toda $\mathbb{P} \in \mathrm{H}_* \wedge \mathrm{H}_0$, vale que

$$\widehat{w}_n \xrightarrow{d} \chi^2(k),$$

onde $\widehat{w}_n \coloneqq n(\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{a})'\big(\boldsymbol{A}\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{A}'\big)^{-1}(\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{a})$ e onde $\chi^2(k)$ denota a distribuição qui-quadrado com k graus de liberdade.

Exercícios

Quaisquer erros nos enunciados são — por hipótese — intencionais.

Exercício 77. Verifique a validade da equação (30).

Exercício 78. Justifique por que a convergência em (32) é válida.

Exercício 79. Mostre que, sob ergodicidade, $n^{-1}\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}} \xrightarrow{p} \mathbb{E}(u_1^2)$.

Exercício 80. Suponha que $\{(y_n, \boldsymbol{x}_n)\}_{n\geq 1}$ é uma sequência iid satisfazendo as seguintes condições:

1. $\mathbb{E}(y_1^4) < \infty$.
2. $\mathbb{E}(|\boldsymbol{x}_1|^r) < \infty$ para algum $r \geq 4$.
3. $\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')$ é positiva definida.

Mostre que

$$\max_{1\leq i\leq n} \big|\widehat{u}_i - u_i\big| = o_{\mathbb{P}}\big(\sqrt[4]{n}/\sqrt{n}\big).$$

Exercício 81. Demonstre o Teorema 33 no caso em que $d = 0$ (uma covariável, sem termo constante). A norma que aparece nos itens 2. e 3. do teorema pode ser tomada como qualquer norma matricial.

Exercício 82. Demonstre o Corolário 34.

Exercício 83. Demonstre o Teorema 36.

Exercício 84. Nas condições do exemplo 35, mostre que

$$\mathbb{P} \in \mathrm{H}_* \wedge (\neg \mathrm{H}_0) \Rightarrow \lim_{n\to\infty} \mathbb{P}\Big(\Big|\sqrt{n}\ \widehat{\sigma}_{kk}^{-1}\big(\widehat{\beta}_{nk} - b\big)\Big| > \xi_\alpha\Big) = 1,$$

onde ξ_α satisfaz $\mathbb{P}(\mathrm{N}(0,1) \leq \xi_\alpha) = 1 - \alpha/2$, $\alpha \in (0,1)$. A função

$$\mathrm{power}(\mathbb{P}) \coloneqq \mathbb{P}\Big(\Big|\sqrt{n}\ \widehat{\sigma}_{kk}^{-1}\big(\widehat{\beta}_{nk} - b\big)\Big| > \xi_\alpha\Big), \quad \mathbb{P} \in \mathrm{H}_* \wedge (\neg \mathrm{H}_0),$$

é dita *poder* do teste.

Exercício 85. Prove o seguinte resultado:

Teorema 37. Suponha que $\{(y_i, \boldsymbol{x}_i)_{i\geq 1}\}$ é uma sequência ergódica satisfazendo as seguintes condições (H_* = ergodicidade∧a∧b∧c∧d):

(a) $\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')$ é invertível.
(b) $\mathbb{E}\big(|u_1^2\boldsymbol{x}_1\boldsymbol{x}_1'|\big) < \infty$.
(c) $\mathbb{E}\big(\boldsymbol{x}_{\ell+1}u_{\ell+1} \,\big|\, (\boldsymbol{x}_i, u_i), 1 \leq i \leq \ell\big) = \mathbf{0}$ para todo $\ell \geq 1$.
(d) $\mathbb{E}(u_1^2\boldsymbol{x}_1\boldsymbol{x}_1')$ é invertível,

onde $u_i \coloneqq y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, com $\boldsymbol{\beta} = \big(\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1')\big)^{-1}\mathbb{E}(\boldsymbol{x}_1 y_1)$. Seja ainda $h\colon \mathbb{R}^{d+1} \to \mathbb{R}^k$ uma função cujas derivadas parciais existem e tal que a matriz de derivadas parciais $\partial h_{\boldsymbol{\beta}}$ satisfaça $\mathrm{rank}(\partial h_{\boldsymbol{\beta}}) = k$. Considere a hipótese nula $\mathrm{H}_0 = \{\mathbb{P}\colon h(\boldsymbol{\beta}) = \mathbf{0}\}$. Então, para toda $\mathbb{P} \in \mathrm{H}_* \wedge \mathrm{H}_0$, vale que

$$\widehat{w}_n \xrightarrow{d} \chi^2(k),$$

onde $\widehat{w}_n \coloneqq nh(\widehat{\boldsymbol{\beta}})'\big(\partial h_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\Sigma}}_n\partial h_{\widehat{\boldsymbol{\beta}}}'\big)^{-1}h(\widehat{\boldsymbol{\beta}})$ e onde $\chi^2(k)$ denota a distribuição qui-quadrado com k graus de liberdade.

Exercício 86. Proponha um intervalo de confiança para β_k. Dê os detalhes, em particular indicando as propriedades assintóticas do intervalo proposto. Você conseguiria propor uma *região* de confiança para $\boldsymbol{\beta}$? E um intervalo de confiança (condicional) para $\boldsymbol{\xi}'\boldsymbol{\beta}$?

Referências Bibliográficas

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 3rd edition, 1995.

Olav Bjerkholt. On the Founding of the Econometric Society. *Journal of the History of Economic Thought*, 39(2), 2017.

George Box. Robustness in the Strategy of Scientific Model Building. In Robert L. Launer and Graham N. Wilkinson, editor, *Robustness in Statistics*, pages 201 – 236. Academic Press, 1979.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

John Geweke, Joel L. Horowitz, and Hashem Pesaran. *Econometrics*, pages 1–44. Palgrave Macmillan UK, London, 2017.

Trygve Haavelmo. The Probability Approach in Econometrics. *Econometrica*, 12:iii–115, 1944.

Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

David F. Hendry. Econometrics – Alchemy or Science? *Economica*, 47(188):387–406, 1980.

Kevin D. Hoover. The Methodology of Econometrics. In Terence C. Mills and Kerry Patterson, editors, *Palgrave Handbook of Econometrics*, volume 1. Palgrave Macmillan UK, 2006.

Joseph B. Kadane. *Principles of Uncertainty*. Chapman and Hall/CRC, 2011.

Samuel Karlin and Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press, Boston, 2nd edition, 1975.

R.L. Plackett. The discovery of the method of least squares. *Biometrika*, 59(2), 1972.

Juliet Popper Shaffer. The Gauss-Markov Theorem and random regressors. *The American Statistician*, 45(4):269–273, 1991.

John Stachurski. *A primer in econometric theory*. MIT Press, 2016.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Univ. Press, 1998.

David Williams. *Weighing the Odds: A Course in Probability and Statistics*. Cambridge University Press, 2001.