Notas de aula de econometria III

Daniel D. Santos

Agosto/ 2009

1 Parte 1: Concepção de um exercício empírico

> "Estimators should be selected on the basis of their ability to answer well-posed economic problems with minimal assumptions" (J. Marschak)

1.1 Aula 1: Causalidade

Genericamente, podemos dividir em duas as finalidades de um exercício econométrico que vise responder a uma pergunta econômica. Em primeiro lugar, nosso interesse pode estar em identificar e mensurar uma relação de **causalidade** entre duas variáveis. Nesse espírito, incluem-se tanto testes de validação de teorias econômicas, que por exemplo propõe diretamente uma relação de causa-efeito que gostaríamos de saber se é válida; quanto avaliações de políticas públicas e projetos empresariais, onde o objetivo é verificar se determinado comportamento observado em uma variável de resultado foi efetivamente *causado* pela política ou projeto em questão. Em segundo lugar, nosso interesse pode estar focado em **prever** um valor futuro de um indicador, com base no co-movimento observado até o presente entre defasagens deste indicador e de outras variáveis. Neste curso, exploraremos a primeira finalidade com mais afinco.

1.1.1 Relações determinísticas

Considere primeiro uma relação de causalidade do tipo:

$$y = \mu(x)$$

Em princípio, há tanto situações em que primeiro formulamos uma teoria que explica a relação entre y e x através da função $\mu(.)$, quanto situações em que primeiro percebemos que tal relação se manifesta nos dados para depois tentar explicar teoricamente o motivo deste fato. O importante é notar que o efeito causal de uma variação em x sobre y é medido por $\frac{\partial y}{\partial x}$, no caso de uma variação marginal em x, ou $\frac{\Delta y}{\Delta x}$, no caso de uma variação discreta. Como estimaríamos esta relação se nos fosse dada uma base com informações sobre y e x para uma amostra de indivíduos? Uma idéia seria simplesmente a de escolher um indivíduo para cada valor de x, e computar o respectivo valor de y ao longo do domínio de x. Outra idéia seria repartir a amostra em grupos com diferentes valores de x (se X for idade, seriam o grupo dos que têm 0 anos, os do que têm 1 ano de idade, etc. Se X for sexo, os grupos seriam homens e mulheres, e assim por diante) e tomar a média de y em cada grupo. Como nossa função μ é determinística, a média de y na célula homogênea em $X = x$ seria

$$\overline{y}|_{X=x} = \frac{1}{N_x}\sum_{i=1}^{N_x}\mu(x) = \frac{N_x}{N_x}\mu(x) = \mu(x)$$

O efeito de um aumento de X de, digamos, x para x*, é então medido por:

$$\frac{\Delta y}{\Delta x} = \frac{\mu(x^*) - \mu(x)}{x^* - x}$$

Finalmente, note que no caso particular em que $\mu(x) = a + bx$, temos que

$$\frac{\Delta y}{\Delta x} = \frac{\partial y}{\partial x} = b$$

O passo seguinte é considerar uma função, ainda determinística, de 2 variáveis:

$$y = \mu(x, z)$$

Agora, temos que definir qual efeito causal desejamos estimar. Uma possibilidade é o efeito direto (ou Marshalliano): $\frac{\partial y}{\partial x}(X = x, Z = z)$. Qual a interpretação? Este efeito mede, para um indivíduo que possua $(X = x, Z = z)$, o que acontece com seu valor de y se aumentarmos marginalmente x, *mantendo z constante*. Usualmente, é essa a primeira noção de efeito causal que temos em mente.

Outra possibilidade é medir o efeito total $\frac{dy}{dx} = \frac{\partial y}{\partial x}(X = x, Z = z) + \frac{\partial y}{\partial \varepsilon}(X = x, Z = z)\frac{\partial \varepsilon}{\partial x}(X = x, Z = z)$. Nesse caso, estamos incorporando tanto o efeito direto quanto o efeito que uma mudança de x pode ter sobre y por forçar uma modificação de z. Em alguns casos, pode ser este o parâmetro que nos dará a resposta à pergunta que temos em mente.

No próximo passo, note que se fizermos a hipótese de *separabilidade aditiva*:

$$y = \mu(x, z) = \mu(x) + f(z)$$

então teremos que $\frac{\partial y}{\partial x}(X = x, Z = z) = \frac{\partial y}{\partial x}(Z = z)$ e $\frac{\partial y}{\partial \varepsilon}(X = x, Z = z) = \frac{\partial y}{\partial \varepsilon}(Z = z)$. Se além disso supusermos que $\mu(x) = a + bx$, então $\frac{\partial y}{\partial x}(X = x) = b$. Assim como antes, se observarmos y, x e z, basta formar células homogêneas em x e z, e calcular o valor de y em cada uma delas. Vamos definir a partir dessa formulação $\varepsilon = f(z)$.

Agora, se fizermos um gráfico tridimensional com y no eixo das ordenadas e x e ε nas abcissas, teremos uma superfície determinística, mas se fizermos um gráfico em $\Re^2$ com y nas ordenadas e x nas abcissas, teremos uma nuvem de pontos, onde para cada valor de $X = x$, a distribuição de $Y|X = x$ segue distribuição parecida com a de $\varepsilon|X = x$, apenas deslocada por $\mu(x)$. Por outro lado, continuo podendo calcular o efeito direto de x sobre y, $\frac{\Delta y}{\Delta x} = \frac{\mu(x^*)-\mu(x)}{x^*-x}$, já que posso simplesmente escolher dois indivíduos com o mesmo valor de z mas x diferentes, e fazer $\frac{\Delta y}{\Delta x} = \frac{[\mu(x^*)+f(z)]-[\mu(x)+f(z)]}{x^*-x}$

Finalmente, note que se for verdade que $E[\varepsilon|X] = 0$, para qualquer $X = x$, então

$$E[y|X = x] = \mu(x)$$

o que significa que se computarmos simplesmente a média de y em cada célula homogênea em $X = x$, teremos uma medida de $\mu(x)$, e portanto tanto $\mu(x)$ quanto $\frac{\Delta y}{\Delta x} = \frac{\mu(x^*)-\mu(x)}{x^*-x}$ podem ser computados[1].

Uma teoria que conclua que y, x e z devem estar deterministicamente relacionados apresenta uma implicação bastante forte sobre os dados: basta que haja um único caso em que dois indivíduos com mesmo (x,z) possuam y's distintos para refutar a teoria. Eis porque raramente este caso aparece em problemas empíricos econômicos.

[1] Estou sendo propositalmente informal nessa explicação. Se por exemplo X for uma variável contínua, não é possível construir uma célula de indivíduos com o mesmo $X = x$, mas no entanto continua sendo verdade que a função μ pode ser genericamente estimada se supusermos que o modelo é aditivamente separável e que $E[\varepsilon|X] = 0$ (exogeneidade estrita).

1.1.2 Relações estocásticas

A pergunta que nos fazemos agora é: o que acontece se ε não for observado? Mais à frente discutiremos outras possíveis interpretações para a presença de um termo estocástico (aleatório) em uma relação econômica, mas inicialmente suponha que ε é simplesmente um determinante não-observado de y. Como mostramos acima, mesmo que tivéssemos em nossa base de dados y, x e ε, ainda assim poderíamos estimar o efeito direto de x sobre y usando apenas informações sobre x e y, caso nosso modelo fosse aditivamente separável e x fosse estritamente exógeno com respeito a ε.

1.2 Linearidade e Mínimos Quadrados Ordinários

Vamos a partir de agora estudar um pouco mais o caso em que $\mu(x) = a + bx$ (hipótese de *linearidade*). Veremos mais adiante várias estratégias para chegar ao estimados de mínimos quadrados ordinários (MQO). O que importa agora é mostrar que ele funciona neste caso e porque funciona.

Em uma relação do tipo:

$$\begin{aligned} y_i &= \alpha + \beta x_i + \varepsilon_i \\ E(\varepsilon|X) &= 0 \end{aligned}$$

definimos o estimador de MQO $\left(\widehat{\alpha}, \widehat{\beta}\right)$ de (α, β) em uma amostra $\left(Y^N, X^N\right) = (y_1, ..., y_N, x_1, ..., x_N)$, como sendo:

$$\begin{aligned}\widehat{\alpha} &= \overline{y} - \widehat{\beta}\overline{x} \\ \widehat{\beta} &= \frac{\widehat{cov}(x,y)}{\widehat{var}(x)} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})\, y_i}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})\, x_i} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})\, y_i}{\sum_{i=1}^{N}(x_i - \overline{x})\, x_i}\end{aligned}$$

onde $\overline{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$ denota a média amostral da variável aleatória Z.

Exercise 1 *Mostre que para quaisquer duas variáveis aleatórias z,w:* $\sum_{i=1}^{N}(z_i - \overline{z})\, w_i = \sum_{i=1}^{N}(z_i - \overline{z})(w_i - \overline{w})$

Exercise 2 *Mostre que para qualquer variável aleatória z e constante A:* $var(Az) = A^2 var(z)$

Exercise 3 *Mostre que para qualquer variável aleatória z e constante A:* $cov(Az) = 0$

Exercise 4 *Mostre que para qualquer variável aleatória z e constante A:* $var(A + z) = var(z)$

Exercise 5 *Mostre que para quaisquer duas variáveis aleatórias z,w:* $var(z + w) = var(z) + var(w) + 2cov(z, w)$

Exercise 6 *Mostre que para quaisquer três variáveis aleatórias z,w,y:* $cov(y + z, w) = cov(y, w) + cov(z, w)$

Neste contexto:

$$\begin{aligned}\widehat{\beta} &= \frac{\widehat{cov}(x, \alpha + \beta x + \varepsilon)}{\widehat{var}(x)} \\ &= 0 + \beta + \frac{\widehat{cov}(x,\varepsilon)}{\widehat{var}(x)} \\ &= \beta + \frac{\widehat{cov}(x,\varepsilon)}{\widehat{var}(x)} \\ P\lim\left(\widehat{\beta}\right) &= \beta + \frac{cov(x,\varepsilon)}{var(x)}\end{aligned}$$

Dada a hipótese de exogeneidade estrita, $E(\varepsilon|X) = 0$, temos que $cov(x, \varepsilon) = 0$, pois

$$\begin{aligned} cov\left(x,\varepsilon\right) &= E\left(x\varepsilon\right) - E\left(x\right)E\left(\varepsilon\right) \\ &= E\left[xE\left(\varepsilon|x\right)\right] - E\left(x\right)E\left(E\left(\varepsilon|X\right)\right) \\ &= E\left[x0\right] - E\left(x\right)0 = 0 \end{aligned}$$

Nesse argumento, usei a Lei das Expectativas Iteradas, que estabelece que, para quaisquer duas variáveis aleatórias z,w: $E\left(z\right) = E\left[E\left(z|W\right)\right]$ [2].

De fato, com isso não apenas demonstramos que $P\lim\left(\widehat{\beta}\right) = \beta$, como também que exogeneidade estrita é mais do que precisamos para estimar o coeficiente $\widehat{\beta}$ por MQO, pois $cov\left(x,\varepsilon\right) = 0$ seria suficiente e é implicada por exogeneidade estrita.

Outro ponto interessante é notar que o princípio que faz com que MQO funcione num contexto linear em que (x,ε) são não-correlacionadas é o mesmo que faz com que a omissão de um regressor z (observável) em uma regressão linear não afete a consistência de $\widehat{\beta}$ caso $cov(x,z) = 0$. Para ver isso, suponha que o modelo verdadeiro seja

$$\begin{aligned} y_i &= \alpha + \beta x_i + \gamma z + \varepsilon_i \\ E\left(\varepsilon|X,Z\right) &= 0 \end{aligned}$$

mas que em nossa regressão z esteja omitido. Neste caso:

[2] Derivamos essa demonstração em sala, mas considere o seguinte exemplo: Suponha que você quer calcular a média de altura (z) da sala. Uma forma de fazê-lo é medir a altura de cada aluno e tomar a média, E(z). Outra forma é tomar primeiro a média de homens e mulheres (chame a variável sexo de w) separadamente, E(z|w=homem) e E(z|w=mulher), e depois calcular a média de altura da sala computando uma média ponderada entre essas duas médias. Note que E(z|W) é uma função de w! (para cada valor de W=w tenho uma média diferente). Neste caso, os ponderadores devem ser as respectivas frequências de homens e mulheres na sala (distribuição f(w)), e temos E[E(z|W)].

$$\begin{aligned}\widehat{\beta} &= \frac{\widehat{cov}\left(x, \alpha + \beta x + \gamma z + \varepsilon\right)}{\widehat{var}\left(x\right)} \\ &= \beta + \gamma \frac{\widehat{cov}\left(x, z\right)}{\widehat{var}\left(x\right)} + \frac{\widehat{cov}\left(x, \varepsilon\right)}{\widehat{var}\left(x\right)}\end{aligned}$$

e se $cov(x, z) = 0$, então novamente $P\lim\left(\widehat{\beta}\right) = \beta$.

A pergunta seguinte é: Será que MQO estima algum parâmetro interessante caso a relação entre y e x não seja linear?

Inicialmente, vamos considerar um modelo mais geral do tipo:

$$\begin{aligned}y_i &= \mu\left(x_i\right) + \varepsilon_i \\ E\left(\varepsilon | X\right) &= 0\end{aligned}$$

Se usarmos uma expansão de Taylor de primeira ordem (aproximação linear) em torno da esperança de X, x^e, temos que:

$$\mu\left(x\right) = \mu\left(x^e\right) + \left(x - x^e\right)\frac{\partial \mu\left(x^e\right)}{\partial x} + R$$

tal que R é pequeno para valores de x próximos a x^e, ou para funções com pouca curvatura (próximas de serem lineares). A aproximação quando R é pequeno fica:

$$\mu\left(x\right) \approx \mu\left(x^e\right) + \left(x - x^e\right)\frac{\partial \mu\left(x^e\right)}{\partial x}$$

e nesse caso o estimador de MQO é:

$$\begin{aligned}
\widehat{\beta} &= \frac{\widehat{cov}\left(x, \mu\left(x_i\right) + \varepsilon_i\right)}{\widehat{var}\left(x\right)} \\
&\approx \frac{\widehat{cov}\left(x, \mu\left(x^e\right) + \left(x - x^e\right)\frac{\partial \mu(x^e)}{\partial x} + \varepsilon_i\right)}{\widehat{var}\left(x\right)} \\
&= \frac{\widehat{cov}\left(x, \overbrace{\mu\left(x^e\right) + x^e \frac{\partial \mu\left(x^e\right)}{\partial x}}^{\text{constante}}\right)}{\widehat{var}\left(x\right)} + \frac{\widehat{cov}\left(x, x\right)}{\widehat{var}\left(x\right)}\frac{\partial \mu\left(x^e\right)}{\partial x} + \frac{\widehat{cov}\left(x, \varepsilon_i\right)}{\widehat{var}\left(x\right)} \\
&= 0 + \frac{\partial \mu\left(x^e\right)}{\partial x} + 0 = \frac{\partial \mu\left(x^e\right)}{\partial x}
\end{aligned}$$

ou seja, MQO fornece a melhor aproximação linear do efeito marginal avaliado na esperança de X. Tal aproximação é tão melhor quanto menor for a curvatura de $\mu\left(x^e\right)$ e quanto menor for a dispersão de X.

1.2.1 Projeção versus Regressão

Suponha que você tenha uma base de dados com N observações das variáveis Y e X. Independentemente do modelo que realmente explica a relação de Y e X (no exemplo acima, onde $y = \mu\left(x\right) + \varepsilon$, independentemente da forma funcional de μ...), nós sempre podemos ajustar uma reta que maximiza a parcela da variância de y explicada pela variação de x, fazendo:

$$\widehat{y} = \overline{y} + \frac{cov\left(x, y\right)}{var\left(x\right)}\left(x - \overline{x}\right)$$

Note na expressão acima, que $\overline{y}, \overline{x}$ e $\frac{cov(x,y)}{var(x)}$ são números (constantes), e que o lado direito da equação é portanto função de x apenas. Esta função é chamada de *projeção* de y em x, e pode ou não coincidir com a função *regressão* de y em x, definida como sendo $E\left[Y|X\right]$. No

caso em que o modelo verdadeiro estabelece que:

$$\begin{aligned} y_i &= a + bx_i + e_i \\ E[e|X] &= 0 \end{aligned}$$

as funções projeção e regressão coincidem.

Vimos no primeiro teste que a função regressão pode em alguns casos ser trivialmente estimada através de médias condicionais amostrais. No exemplo dado, Y era salário e X escolaridade. Se repartirmos nossa amostra em grupos homogêneos de escolaridade, podemos medir a média salarial dos analfabetos, dos que possuem primário incompleto, e assim por dainte até chegar aos indivíduos com doutorado completo. E isso tudo sem precisar impor qualquer forma funcional à função μ. De fato, essa estratégia é um exemplo do que chamamos de regressão não-paramétrica de Y em X, e apesar do nome complicado não usou nada mais sofisticado do que médias.

A pergunta então é: se podemos estimar de modo bem mais geral a relação entre Y e X, por que insistimos com frequência em impor (na maioria das vezes arbitrariamente) hipóteses de linearidade e parameterizações?

Há três justificativas para esse procedimento não ser estritamente dominado pela regressão não-paramétrica. A primeira é de interpretação. É relativamente fácil visualizar uma curva em um gráfico de salários e escolaridade capturando de forma bastante flexível a função $E[Y|X]$, ainda que o efeito marginal (ou direto) de aumentarmos X de x para x^* varie com o nível de x (no caso linear, o efeito é o mesmo independentemente do valor

de $X = x$ em se considere a variação, mas no caso não-linear isso não é verdade). Com múltiplos regressores, a análise de uma relação completamente flexível torna-se bem mais difícil de interpretar. O efeito marginal de uma variação em um determinado regressor X_k passa a depender do ponto $(x_1, ..., x_k, ..., x_K)$ em que é avaliado. Perguntas do tipo "o que aconteceria com a renda média da população se elevássemos a média de escolaridade em 1 ano?" agora têm múltiplas respostas. Por exemplo, nesse caso faz toda diferença se a média de educação for elevada educando um pouco mais os que já tem ensino médio completo ou dando educação aos analfabetos, se esses aumentos são focados no sudeste ou no nordeste, em jovens ou adultos, etc (quantos sejam os determinantes de salário). Formas mais flexíveis dão respostas mais precisas, mas demandam mais tempo para interpretar e transmitir.

A segunda justificativa relaciona-se ao fato de que nem sempre nossos regressores são discretos. Suponha que ao invés de escolaridade o regressor considerado fosse idade. Em princípio, não se pode encontrar dois indivíduos que tenham nascido exatamente no mesmo instante de tempo, de modo que agrupar a amostra em indivíduos com rigorosamente a mesma idade para depois tomar a média salarial em cada grupo deixa de ser algo factível. Hoje já existem técnicas para aproximar arbitrariamente bem uma regressão não-paramétrica com regressores contínuos, mas tais técnicas exigem elevada capacidade computacional e ainda são inviáveis se o número de regressores for grande.

A terceira justificativa é que em geral maior flexibilidade consome mais graus de liberdade para que se possa estimar a regressão. Um exemplo comum em livros-texto de estatística é considerar o caso em que nossa base contém N observações e nosso modelo contém N-1

regressores linearmente independentes. Neste caso, é possível mostrar que o ajuste do modelo é perfeito (todos os resíduos se igualam a zero, $R^2 = 1$, etc.). O caso acima é análogo ao de se ajustar uma reta a dois pontos quaisquer, ou um polinômio quadrático a três pontos, etc. O problema neste caso é a insuficiência de graus de liberdade para estimar todos os parâmetros do modelo, e a principal consequência é que o modelo deixa de ter validade externa à amostra em que foi estimado. Como consequência, se a relação entre Y e X for verdadeiramente linear e resolvermos estimá-la de modo mais flexível, nossa estimação será menos precisa e a capacidade de extrapolar suas conclusões para casos fora da amostra, diminuída.

1.2.2 Por quê tanto interesse na função regressão?

Se estivermos dispostos a supor que (i) $y = \mu(x) + \varepsilon$ e (ii) $E(\varepsilon|X) = 0$, nosso interesse principal estará em estimar o que aconteceria com y se elevássemos X de x para x^*. Em outras palavras, as perguntas que nos interessam passam pela estimação de

$$\Delta = \frac{\mu(x^*) - \mu(x)}{x^* - x}$$

Se nossas hipóteses forem verdadeiras, então $E(Y|x) = \mu(x)$, com a vantagem que $E(Y|x)$ é um momento amostral, que pode portanto ser diretamente estimado a partir de informações sobre X e Y.

2 Conteúdo empírico de um modelo

Uma teoria costuma ser um conjunto de argumentos logicamente encadeados, mas que pode ou não possuir conteúdo empírico, isto é, implicações que tenham como contrapartida limi-

tações à realidade, possíveis de serem testadas. Se uma teoria não possui conteúdo empírico, dizemos que é tautológica, e suas conclusões não podem ser contestadas por elementos observados na realidade. Um exemplo disso é uma teoria que busque explicar o comportamento decisório dos consumidores baseada apenas nos axiomas fundamentais de completude, transitividade e reflexividade. Isso porque para qualquer comportamento decisório observado nos dados podemos encontrar preferências que atendam a estes axiomas e que racionalizem tais decisões. Quando uma teoria possui conteúdo empírico, é em geral possível conceber abstratamente ao menos um exercício empírico que permita, com o auxílio de dados apropriados, verificar se as limitações impostas à realidade pela teoria de fato se verificam ou não.

Uma pergunta empírica corretamente formulada tem por tras um modelo teórico que necessariamente impõe restrições à realidade. Se por exemplo dizemos que a forma como duas variáveis Y e X se relacionam é através da função $y = a + bx$, podemos não apenas escolher os parâmetros a e b que melhor adequam a realidade à teoria (e que seriam portanto os parâmetros verdadeiros caso a teoria fosse correta), como também testar se neste caso, mais favorável à teoria, as restrições por ela impostas são coerentes com o comportamento observado nos dados.

Como discutimos anteriormente, um modelo que implique restrições determinísticas entre X e Y pode ser trivialmente estimado e testado a partir de observações de X e Y. No exemplo acima, existe no máximo um par de parâmetros (a,b) que pode satisfazer a equação $y = a + bx$ para todos os pares observados de (y, x). Se tais relações forem estocásticas,

digamos $y = a + bx + e$, onde apenas x e y sejam observados, para cada par (a,b) escolhido há uma chance de que o modelo seja verdadeiro ainda que a escolha de (a,b) não pareça explicar satisfatoriamente a relação entre y e x, ou de que o modelo esteja errado ainda que o par (a,b) eleito pareça de fato explicar tal relação. Ainda que sempre paire essa incerteza, algum par (a,b) será eventualmente escolhido através do estimador proposto, que invariavelmente utilizará as restrições à realidade impostas pela teoria para selecionar o par que maximize as chances de (a,b) ser o verdadeiro par, caso a teoria esteja correta.

Considere agora um modelo caracterizado por um conjunto de parâmetros Θ resultante de uma teoria que implica que R restrições sobre a realidade sejam válidas. Dizemos que o modelo é unicamente identificado (ou simplesmente identificado) se houver um único vetor de parâmetros Θ^* que, no conjunto de modelos candidatos a satisfazer R, possua chances de ser o que de fato racionaliza a realidade maiores do que qualquer concorrente $\Theta^{**} \neq \Theta^*$. Evidentemente, as dimensões de Θ e R importam para saber se um modelo é unicamente identificado ou não. Se nosso modelo tiver muitos parâmetros para estimar mas a teoria implicar em poucas restrições sobre a realidade, a chance de que dois modelos caracterizados por Θ^* e Θ^{**} expliquem igualmente bem a realidade, aumentam. Se por exemplo a teoria sugerir que $y = a + bx$, mas houver apenas um par observado (x,y), é fácil ver que há um contínuo de pares (a,b) que racionaliza igualmente bem o par observado. Com mais pares observados, obtemos mais restrições, pois agora o mesmo par (a,b) passa a ter que explicar as observações $(x_1, y_1), (x_2, y_2)$, ... Por outro lado se $y = a + bx + e$ e $y^2 = a\cos(x) + u$, temos novamente que o mesmo par (a,b) tem que racionalizar mais de uma restrição sobre a

realidade, e as chances de que dois pares distintos (a^*, b^*) e (a^{**}, b^{**}) racionalizem igualmente bem os mesmos dados, diminuem. Um modelo associado a uma teoria que imponha mais restrições sobre a realidade do que o mínimo necessário para identificar unicamente Θ é dito super-identificado, e um que possua menos restrições do que o mínimo necessário para identificar Θ é dito sub-identificado. É possível, por exemplo, que um modelo seja sub-identificado em uma base de dados e super-identificado em outra base de dados.

3 Estratégias de estimação

Uma vez que um modelo proposto deva satisfazer a um determinado conjunto de restrições, estratégias de estimação são elaboradas de modo a escolher elementos do modelo (no nosso caso, parâmetros) que maximizem a chance de que tais restrições sejam válidas, dada a disponibilidade dos dados existentes.

Tipicamente, um modelo para nós será do tipo $y = M(x, \varepsilon; \Theta)$, ou seja, trata da relação de uma variável (ou vetor de variáveis) dependente y com conjuntos de determinantes observáveis (x) e não-observáveis (ε), sendo ainda caracterizado por um vetor de parâmetros, Θ.

As restrições impostas podem dizer respeito (i) à forma como y se relaciona com x e ε, por exemplo, $y = \mu(x) + \varepsilon$ (separabilidade aditiva) ou $y = a + bx + \varepsilon$ (linearidade); ou (ii) à distribuição conjunta dos determinantes de y, por exemplo $x \underline{\parallel} \varepsilon$ (independência), $x \perp \varepsilon$ (ortogonalidade), $E(\varepsilon|X) = 0$ (exogeneidade estrita), $f(\varepsilon|X) = N(0, \sigma^2)$ (normalidade + homocedasticidade). Note que invariavelmente estas restrições podem ser escritas como lim-

itações ao comportamento do elemento não-observável, ε. Novamente, restrições envolvendo somente y e x são em geral consideradas muito fortes para racionalizar o comportamento humano e facilmente rejeitadas pelos dados.

Em todos os exemplos analisados no curso, ao menos separabilidade aditiva será suposta.

3.1 Minimização de função perda

Na maioria dos exercícios empíricos, ε é uma variável aleatória com média zero. Em modelos lineares, tal resultado emerge sem perda de generalidade, pois se

$$\begin{aligned} y &= a + bx + \varepsilon \\ E(\varepsilon) &= c \end{aligned}$$

, podemos sempre redefinir ε como $\varepsilon^* = \varepsilon - c$, de modo que $E(\varepsilon^*) = 0$ e $y = a^* + bx + \varepsilon^*$, com o intercepto agora sendo $a^* = a - c$ (de fato, o intercepto a que conseguimos estimar é sempre a diferença entre o intercepto verdadeiro da regressão e a média do componente não-observável. Felizmente na maioria dos casos nosso parâmetro de interesse é b, que no caso linear é interpretado como $\partial y/\partial x$, ou seja, o efeito marginal de um aumento de x sobre y, mantidos os outros determinantes de y constantes).

Por outro lado, sempre que elaboramos um modelo teórico para descrever um fenômeno, gostaríamos de poder explicar o máximo deste fenômeno com variáveis que podemos observar e controlar. No caso extremo, se toda a variação em nossa variável dependente y pudesse ser explicada pela variação em determinantes observáveis de y, x, então $\varepsilon = 0$. A estratégia de

minimização de função perda cria uma função $L(\varepsilon)$ que penaliza valores de ε distantes de zero. Para ser mais preciso, L possui as seguintes propriedades:

$L(0) = 0$ - ou seja, se $\varepsilon = 0$ a perda é nula

$L(\varepsilon + \delta) > L(\varepsilon)$ sempre que $\|\varepsilon + \delta\| > \|\varepsilon\|$ - ou seja, sempre que ε se afasta de zero a função perda c

Exemplos de funções que satisfazem tais propriedades são:

$$\text{perda quadrática} : L(\varepsilon) = \varepsilon^2$$

$$\text{perda absoluta} : L(\varepsilon) = |\varepsilon|$$

$$\text{perda absoluta quantílica} : L(\varepsilon; \tau) = [\tau - \mathbf{1}(\varepsilon > 0)]\,\varepsilon$$

Como nossa população possui vários indivíduos, nosso estimador será aquele que minimiza a perda média na população. Como isso funciona?

Primeiro, considere um modelo do tipo $y = \mu(x) + \varepsilon$. De um modo geral, gostaríamos de estimar a função $\mu(.)$, pois o efeito de uma mudança de x para x^* sobre y pode em geral ser aproximado por $\mu(x^*) - \mu(x)$. Se nosso estimador é baseado na minimização da perda média incorrida pela escolha de uma determinada função $\mu(.)$, então o problema pode ser escrito como:

$$\min_{\mu(.)} E[L(\varepsilon)] = \min_{\mu(.)} E[L(y - \mu(x))]$$

A solução para este problema nos exemplos citados é:

perda quadrática: $E(y|X) \rightarrow$ função de regressão

perda absoluta: $M(y|X) \rightarrow$ mediana condicional

perda absoluta quantílica : $Q_\tau(y|X) \rightarrow$ quantil condicional (regressão quantílica)

A contrapartida amostral deste procedimento é:

$$\min_{\mu(.)} \frac{1}{N} \sum_{i=1}^{N} L\left[y_i - \mu\left(x_i\right)\right]$$

O caso mais comum é considerarmos uma função perda quadrática e partirmos da hipótese de que $\mu(x) = \beta_0 + x_1\beta_1 + ... + x_K\beta_K$. Neste caso, o problema fica:

$$\min_{\beta_0,\beta_1,...,\beta_K} \frac{1}{N} \sum_{i=1}^{N} \left[y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K\right]^2$$

e a solução é $\widehat{\beta}_0^{MQO}, \widehat{\beta}_1^{MQO}, ...\widehat{\beta}_K^{MQO}$.

3.2 Máxima verossimilhança

Suponha agora que a restrição imposta pelo modelo seja a de que $y = \beta_0 + x_1\beta_1 + ... + x_K\beta_K + \varepsilon$, e $\varepsilon|X \sim N(0, \sigma^2)$. Considere inicialmente um conjunto de valores quaisquer para os parâmetros, $(\beta_0^*, \beta_1^*, ..., \beta_K^*, \sigma^*)$. A pergunta que se faz aqui é: qual a chance de eu ter sorteado exatamente minha amostra de indivíduos, se minha população for de fato descrita pelo modelo? Em outras palavras, se for verdade que o mundo é descrito por

$$\begin{aligned} y &= \beta_0^* + x_1\beta_1^* + ... + x_K\beta_K^* + \varepsilon \\ \varepsilon|X &\sim N\left(0, \sigma^{*2}\right) \end{aligned}$$

, qual a probabilidade de que eu tenha sorteado exatamente minha amostra $(y_1, ..., y_N, x_{11}, ..., x_{1N}, ..., x_{KN})$?

A resposta vai depender essencialmente dos valores escolhidos $(\beta_0^*, \beta_1^*, ..., \beta_K^*, \sigma^*)$, e nossa estratégia agora consiste de encontrar os valores $\left(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_K, \widehat{\sigma}\right)$ que maximizam a chance de eu ter sorteado a amostra caso o modelo seja verdadeiro. O raciocínio é o de que se a especificação do nosso modelo for correta, a chance de eu ter sorteado a amostra deveria ser alta, e esta estratégia explora essa propriedade.

No restante desta sessão vou descrever os passos usados para construir estimadores a partir deste princípio, e hipóteses auxiliares que nos ajudarão a simplificar a estimação.

Primeiramente, vamos escrever a probabilidade de que nossa amostra $\left(Y^N, X^N\right) = (y_1, ..., y_N, x_{11}, ..., x_{1N}, ..., x_{KN})$ seja sorteada, caso o modelo seja $y = \mu(x; \Theta) + \varepsilon$ e $\varepsilon|X \sim f_\varepsilon(\Theta)$. Θ neste caso é simplesmente o conjunto total de parâmetros que descreve a função-resposta, $\mu(.)$, e a distribuição de ε, $f(.)$.

Se nossa amostra for aleatória (hipótese!), isto é, se para quaisquer dois indivíduos i, j distintos for verdade que $(x_i, \varepsilon_i) \underline{\perp\!\!\!\perp} (x_j, \varepsilon_j)$ ($\underline{\perp\!\!\!\perp}$ representa independência estatística, o que formalmente pode ser escrito como $f(x_i, \varepsilon_i, x_j, \varepsilon_j) = f(x_i, \varepsilon_i) f(x_j, \varepsilon_j)$):

$$\begin{aligned}\Pr\left(Y^N, X^N|\Theta\right) &= \mathcal{L}\left(\Theta\right) \\ &= \prod_{i=1}^{N} \Pr\left(y_i, x_i|\Theta\right) \\ &= \prod_{i=1}^{N} \Pr\left(y_i|x_i;\Theta\right) \Pr\left(x_i|\Theta\right)\end{aligned}$$

A próxima hipótese utilizada é a de que a distribuição marginal de x não depende do conjunto de parâmetros do modelo, isto é, de que $\Pr\left(x_i|\Theta\right) = \Pr\left(x_i\right)$. Esta hipótese é chamada de exogeneidade fraca.

Agora, utilizando o modelo temos que:

$$\begin{aligned}\varepsilon &= y - \mu\left(x;\Theta\right) \\ y - \mu\left(x;\Theta\right)|X &\sim f_\varepsilon\left(\Theta\right) \Longrightarrow y|X \sim f_y\left(\mu\left(x;\Theta\right);\Theta\right)\end{aligned}$$

A relação entre as densidades condicionais de ε e y, expressas na segunda linha, segue do fato de que, para qualquer constante C, $F_\varepsilon\left(C;\Theta\right) = \Pr\left(\varepsilon < C|X;\Theta\right) = \Pr\left(y - \mu\left(x;\Theta\right) < C|X;\Theta\right) = \Pr\left(y < C + \mu\left(x;\Theta\right)|X;\Theta\right) = F_y\left(C + \mu\left(x;\Theta\right)|X;\Theta\right)$. Portanto:

$$\mathcal{L}\left(\Theta\right) = \prod_{i=1}^{N} f_y\left(\mu\left(x;\Theta\right);\Theta\right) \Pr\left(x_i\right)$$

Neste ponto, já poderíamos estabelecer o problema de encontrar estimadores para os parâmetros do modelo, $\widehat{\Theta}$, como sendo: $\max_\Theta \mathcal{L}\left(\Theta\right)$. O problema é que as funções de distribuição podem ser complicadas, fazendo com que encontrar o máximo do produtório acima

não seja tarefa computacionalmente fácil. Se agora supusermos que, para todo par (x, y) na amostra, $f_y(\mu(x;\Theta);\Theta)\Pr(x_i) > 0$ para qualquer possível valor de Θ (o que basicamente diz que qualquer que seja o valor de Θ, todos os indivíduos da amostra tem de fato alguma chance de terem sido amostrados), então podemos utilizar o fato de que Θ que maximiza $\mathcal{L}(\Theta)$ é o mesmo que maximiza $\ln \mathcal{L}(\Theta)$, para formalizar o problema como:

$$\max_{\Theta} \ln\left[\mathcal{L}(\Theta)\right] = \max_{\Theta} \sum_{i=1}^{N} \ln\left[f_y(\mu(x;\Theta);\Theta)\right] + \ln\left[\Pr(x_i)\right]$$

Em que nossas hipóteses simplificaram o problema? Em primeiro lugar, exogeneidade fraca permite simlpesmente desconsiderarmos $\ln[\Pr(x_i)]$ do problema, pois $\sum_{i=1}^{N} \ln[\Pr(x_i)]$ é uma constante que não afeta o problema de maximização. Em segundo lugar, o produtório virou um somatório. Como as condições de primeira ordem do problema são $\partial \ln\left[\mathcal{L}(\Theta)\right]/\partial\Theta = 0$ e o operador diferença é linear, temos a CPO:

$$\sum_{i=1}^{N} \frac{\partial \ln\left[f_y(\mu(x;\Theta);\Theta)\right]}{\partial\Theta} = 0$$

em geral mais fácil de ser computada do que o produto que tínhamos anteriormente (note que devido à hipótese de que a amostra é aleatória, cada componente da soma acima envolve apenas o par (y, x) de um indivíduo da amostra).

Para ilustrar o método, considere o modelo linear com erros homocedásticos normalmente distribuídos:

$$\begin{aligned} y &= \beta_0 + x_1\beta_1 + ... + x_K\beta_K + \varepsilon \\ \varepsilon|X &\sim N\left(0, \sigma^2\right) \end{aligned}$$

Como resultado da estrutura suposta, temos que

$$y|X \sim N\left(\beta_0 + x_1\beta_1 + ... + x_K\beta_K; \sigma^2\right)$$

Para relembrar, se uma variável Z distribui-se normalmente com média μ e variância σ^2, então:

$$f_z = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right]$$

o que implica que em nosso caso:

$$f_{y|X=x} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\beta_0 - x_1\beta_1 - ... - x_K\beta_K}{\sigma}\right)^2\right]$$

Nosso estimador de máxima verossimilhança então resolve:

$$\begin{aligned} & \max_{\beta_0,\beta_1,...,\beta_K,\sigma} \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\beta_0 - x_1\beta_1 - ... - x_K\beta_K}{\sigma}\right)^2\right] \\ = & \max_{\beta_0,\beta_1,...,\beta_K,\sigma} \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \ln(\sigma) - \frac{1}{2}\left(\frac{y-\beta_0 - x_1\beta_1 - ... - x_K\beta_K}{\sigma}\right)^2 \end{aligned}$$

Novamente, neste exemplo a solução é $\widehat{\beta}_0^{MQO}, \widehat{\beta}_1^{MQO}, ...\widehat{\beta}_K^{MQO}$. Note que o estimador de $\widehat{\sigma}$ neste caso é

$$\begin{aligned} \widehat{\sigma} &= \frac{1}{N}\sum_{i=1}^{N}\widehat{\varepsilon}_i^2 \\ \widehat{\varepsilon}_i &= y_i - \widehat{\beta}_0 - x_{1i}\widehat{\beta}_1 - ... - x_{Ki}\widehat{\beta}_K \end{aligned}$$

que é viesado! Como exercício, mostre que $E(\widehat{\sigma}) = \frac{N-K-1}{N}\sigma$.

3.3 Método dos Momentos

Em grande parte das situações de interesse, as hipóteses implicadas pelo modelo teórico restringem explicitamente momentos da distribuição conjunta de X (variáveis observáveis) e ε (não-observáveis). Exemplos de momentos desta distribuição são:

- $E(\varepsilon) = 0$ (média incondicional de ε igual a zero na população)
- $E(\varepsilon x) = 0$ (ortogonalidade. Média do produto de x e ε igual a zero na população, para todo valor de $X = x$).
- $E(\varepsilon^2) - E^2(\varepsilon) = var(\varepsilon) = \sigma^2$ (homocedasticidade)
- $E(\varepsilon|X) = 0$ (exogeneidade estrita).

O método dos momentos consiste em encontrar um conjunto de parâmetros/ funções que satisfaça simultaneamente uma coleção de momentos amostrais correspondentes aos momentos populacionais implicados pela teoria. Se por exemplo a teoria sugere que $E(\varepsilon) = 0$, e o modelo é descrito pela equação $y = \mu(x) + \varepsilon$, então a contrapartida empírica deste momento em uma amostra aleatória é:

$$\frac{1}{N}\sum_{i=1}^{N}[y_i - \mu(x_i)] = 0$$

Se $\mu(x) = a + xb$:

$$\frac{1}{N}\sum_{i=1}^{N}[y_i - a - x_i b] = 0$$

Outro exemplo seria a condição de ortogonalidade $E(\varepsilon x) = 0$, que teria como contrapartida empírica:

$$\frac{1}{N}\sum_{i=1}^{N} x_i \left[y_i - \mu(x_i)\right] = 0$$

ou, no caso linear:

$$\frac{1}{N}\sum_{i=1}^{N} x_i \left[y_i - a - x_i b\right] = 0$$

Dada uma coleção de momentos, o método consiste simplesmente em encontrar o conjunto de parâmetros que resolve o sistema. Para ilustrar, considere uma regressão linear simples em que valham as seguintes hipóteses:

$$\begin{aligned} y &= a + bx + \varepsilon \\ E(\varepsilon) &= 0 \\ E(\varepsilon x) &= 0 \end{aligned}$$

Neste caso, o sistema de equações correspondente aos momentos amostrais propostos seria:

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \left[y_i - \widehat{a} - x_i\widehat{b}\right] &= 0 \\ \frac{1}{N}\sum_{i=1}^{N} x_i \left[y_i - \widehat{a} - x_i\widehat{b}\right] &= 0 \end{aligned}$$

Da primeira equação, temos que:

$$\begin{aligned}\widehat{a} &= \frac{1}{N}\sum_{i=1}^{N} y_i - \widehat{b}\frac{1}{N}\sum_{i=1}^{N} x_i \\ &= \overline{y} - \widehat{b}\overline{x}\end{aligned}$$

e da segunda:

$$\frac{1}{N}\sum_{i=1}^{N} x_i \left[y_i - \left(\overline{y} - \widehat{b}\overline{x}\right) - x_i\widehat{b}\right] = 0$$

ou

$$\frac{1}{N}\sum_{i=1}^{N} x_i \left[(y_i - \overline{y}) - \widehat{b}\,(x_i - \overline{x})\right] = 0$$

$$\widehat{b} = \frac{\sum_{i=1}^{N} x_i\,(y_i - \overline{y})}{\sum_{i=1}^{N} x_i\,(x_i - \overline{x})}$$

ou seja, MQO.

As demostrações de que este método funciona estão enraizadas na Lei dos Grandes Números (que diz que os momentos amostrais convergem para os populacionais conforme o tamanho da amostra cresça), e no Teorema do Limite Central (que diz que médias de variáveis x definidas sobre um suporte real distribuem-se segundo uma normal com média $E(x)$ e variância $V(x)/N$).

O problema que frequentemente surge quando tentamos aplicar este método é que a teoria pode implicar mais condições de momento do que o mínimo necessário para estimar os

parâmetros de interesse. Se por exemplo a teoria diz que $E(\varepsilon|X) = 0$, esta restrição implica que

$$\begin{aligned}
E(\varepsilon) &= E[E(\varepsilon|X)] = E(0) = 0 \\
E(\varepsilon x) &= E[E(\varepsilon|X = x)\, x] = E(0.x) = 0 \\
E\left(\varepsilon x^2\right) &= E\left[E(\varepsilon|X = x)\, x^2\right] = E\left(0.x^2\right) = 0 \\
&\vdots \\
E(\varepsilon g(x)) &= E[E(\varepsilon|X = x)\, g(x)] = E(0.g(x)) = 0
\end{aligned}$$

ou seja, ε será ortogonal a qualquer função de x, dando margem a uma infinidade de momentos que seriam válidos caso o modelo seja verdadeiro. Vimos por exemplo que no caso da regressão linear os dois primeiros momentos acima são suficientes para estimar $\widehat{a}$ e $\widehat{b}$. A pergunta então é: o que fazer com os demais momentos?

Primeiramente, uma opção é simplesmente ignorá-los. Os estimadores obtidos utilizando apenas os $E(\varepsilon) = 0$ e $E(\varepsilon x) = 0$ serão consistentes. Eventualmente, é possível que não sejam os de menor variância dentre as possibilidades que podem ser construídas a partir da utilização de outros momentos. Eventualmente ainda, é possível que as perguntas que buscamos digam respeito ao comportamento de εx^2 ou $\varepsilon g(x)$, e que para ajustar o modelo a estes momentos uma coleção diferente de momentos amostrais seja mais indicada do que $E(\varepsilon) = 0$ e $E(\varepsilon x) = 0$.

Nos casos em que os momentos populacionais implicados pela teoria são mais do que o conjunto mínimo necessário à estimação dos parâmetros, dizemos que o modelo é *sobre-identificado*. Ainda assim podemos usar os momentos adicionais em nossa estimação.

Primeiramente, defina como M^h o momento $E\left[f^h(\varepsilon)\, g^h(x)\right] = 0$. Esta notação engloba tanto os momentos implicados pela exogeneidade estrita, $E(\varepsilon|X) = 0$, que implica que $E(\varepsilon g(x)) = 0$ (neste caso f^h é a função identidade e $g^h = g$), quanto momentos como homocedasticidade ($E(\varepsilon^2 - \sigma^2) = 0$, onde $f^h = \varepsilon^2 - \sigma^2$ e g^h é a função identidade).

Agora, defina o erro amostral na estimação de M^h, como $e_h = \widehat{M}^h - M^h$. Por definição, construímos um sistema em que todos os momentos são nulos, de modo que $e_h = \widehat{M}^h$, mas é útil manter a notação para clarear o espírito do método (a teoria diz que o momento populacional deveria ser nulo, mas como observamos apenas uma amostra, podemos justificar o fato de que eventualmente o momento amostral não seja nulo. Se observássemos a população inteira, e portanto não tivéssemos o erro amostral como justificativa para o fato de que eventualmente nem todos os momentos estimados vão ser nulos, o modelo seria imediatamente rejeitado. Por outro lado, um sistema sobre-identificado, com mais equações que incógnitas, dificilmente encontraria uma solução que satisfizesse a todas as equações, levando a que alguns momentos estimados inevitavelmente não fossem nulos). Aqui não devemos confundir o erro amostral na estimação do momento h com o determinante não-observado de y, ε.

O terceiro passo é construir um estimador que escolha os parâmetros que satisfaçam ao máximo as restrições do modelo. Idealmente, gostaríamos que todos os e_h fossem iguais a zero, pois a teoria diz que $M^h = 0$. Como isso não é possivel, vamos construir uma função

que penalize desvios de e_h de zero. A forma mais simples de fazê-lo é:

$$\min_{\Theta} \lambda_1 e_1^2 + \lambda_2 e_2^2 + ... + \lambda_h e_h^2 + ... + \lambda_H e_H^2$$

No problema acima, Θ é o conjunto de parâmetros do modelo ($\beta_0, ..., \beta_K, \sigma$, no contexto de uma regressão linear múltipla, por exemplo). e_h^2 é uma medida de distância, que penaliza crescentemente desvios de e_h. Finalmente, $\lambda_h > 0$ representa a importância que eu atribuo em minha estimação para que um determinado momento amostral h seja de fato próximo de zero. Se $\lambda_h = 1$ para todo h, estou no exemplo acima (exogeneidade estrita + homocedasticidade) dizendo que para mim ter a média amostral de ε próximo de zero é tão importante quanto ter a variância do erro constante para todo x. O procedimento utilizado quando o modelo é sobre-identificado é chamado de Método Generalizado dos Momentos, que tem na formulação acima uma versão simplificada. Para uma dada coleção de momentos, $M^1, ..., M^H$, é possível escolher ponderadores $\lambda_1, ..., \lambda_H$ que minimizem a variância do estimador ou que maximizem o ajuste aos dados em alguma dimensão específica do modelo. A consistência dos estimadores, contudo, independe da escolha de λ.

4 Interpretação de ε

Começamos nossa discussão mostrando que uma teoria que implique em relação determinística entre duas variáveis, y e x, é trivialmente estimável e testável. Se $y = \mu(x)$, posso simplesmente ordenar os indivíduos segundo o valor de x e medir para cada um o y correspondente, traçando deste modo a função μ. Neste caso, basta um caso de dois indivíduos

com mesmo x e diferentes y para eu rejeitar minha teoria.

O uso de análise estatística se faz necessário justamente quando há um elemento aleatório no modelo, e vimos também que a forma de estimar os objetos de interesse do modelo depende essencialmente das restrições que o próprio modelo teórico impõe à relação entre a variável dependente y e seus determinantes (x, ε), bem como à distribuição conjunta de x e ε, $f(x, \varepsilon)$.

Nesta seção, o objetivo é discutir as interpretações mais comuns à presença de ε em um modelo econométrico. Essencialmente, as hipóteses que estaremos dispostos a fazer sobre a relação entre y e (x, ε), ou sobre $f(x, \varepsilon)$ dependem fundamentalmente de nossa convicção sobre aquilo que ε de fato representa no modelo. Diferentes interpretações de ε tipicamente conduzem a diferentes hipóteses aceitáveis sobre o comportamento de ε.

4.1 $\varepsilon =$ "sorte"

Uma forma de interpretar ε em uma equação do tipo $y = \mu(x) + \varepsilon$, é a de que a relação entre y e x seria sistematicamente perturbada pela ocorrência de eventos exógenos a esta relação, ocorridos de modo aleatório entre indivíduos e ao longo do tempo.

Neste caso, se por exemplo y for salário e x for escolaridade, a interpretação de ε seria por exemplo a sorte que alguns indivíduos tiveram e outros não de encontrar patrões generosos no momento em que foram contratados.

Se for esta a interpretação de ε, é plausível supor que ε seja independente de x, pois se pessoas com algum tipo de característica tivessem sistematicamente mais "sorte" do que outras, o mais provável é que ε não fosse exatamente sorte e sim algum tipo de remuneração adicional paga aos indivíduos portadores de determinadas características.

Como vimos anteriormente, independência entre ε e x restringe dramaticamente $f(x, \varepsilon)$. Em particular, implica que $E(\varepsilon|X)$ é constante, sendo ainda mais forte. Tal restrição pode ser usada diretamente para formular estimadores.

O problema é que por uma série de razões, esta interpretação raramente é a preferida dos econometristas. A primeira delas é que "sorte" frequentemente viola pressupostos e conclusões comuns da teoria de escolha racional que embasa a economia. É difícil, por exemplo, reconciliar o fato de que dois indivíduos igualmente produtivos estejam recebendo salários distintos devido à "sorte", tanto porque o patrão supostamente mais generoso poderia estar auferindo lucros maiores caso demitisse o atual empregado que recebe mais e contratasse o outro que aceita fazer o mesmo serviço por uma féria menor, quanto porque o empregado que recebe menos poderia em princípio oferecer seus préstimos ao patrão mais generoso e dificilmente se manteria no emprego com menores salários.

A segunda razão é que não aprendemos nada com a "sorte". Se nossa teoria evolui e descobrimos que o modelo anterior estava errado, digamos, porque deixava de incluir aspectos importantes da realidade, não há como aproveitar as estimações feitas até então para descobrir se as novas teorias são de fato as que nos colocam na pista correta. Como veremos adiante, quando ε pode ser interpretado como um determinante não-observável de y (variável omitida), é possível especular sobre a forma como sua exclusão do exercício estaria viesando as estimativas dos demais parãmetros. Se em algum momento se descobrisse uma forma de medir ε (ou parte dele) e se verificasse que os demais parâmetros variaram na direção desejada após a inclusão dessa medição, isto seria indicativo de que de fato foi importante ter

descobrido uma forma de medi-lo, e pode apontar para futuros passos rumo à confirmação da teoria.

A terceira razão é que nos problemas econômicos que estaremos interessados no curso, frequentemente uma parte considerável da variação de y não é explicada pela variação de x. Nestes casos, se ε fosse de fato "sorte", seria de se perguntar qual a utilidade da própria teoria, já que ao fim e ao cabo parte substancial da determinação de y ocorrerá devido a elementos sobre os quais não temos nenhum controle.

4.2 ε = erro de medida

Outra interpretação comum para o termo aleatório em um exercício empírico é o de que este represente um erro na mensuração da variável dependente. Na forma mais simples, diz-se que nosso modelo verdadeiro é por exemplo:

$$y = \mu(x)$$

mas ao invés de observarmos y, observamos apenas $y^* = y - \varepsilon$. Na maioria dos casos é plausível supor que se ε é puro erro de medida, seu valor provavelmente independe do valor de x. Incluindo um intercepto em $\mu(x)$, podemos novamente supor que ε tem média zero. Desse modo, o modelo estimável será

$$\begin{aligned} y^* &= \mu(x) + \varepsilon \\ E(\varepsilon|X) &= 0 \end{aligned}$$

Ainda que seja uma possibilidade, essa interpretação está sujeita aos dois últimos comentários descritos na subseção anterior. Em alguns casos é possível ainda que apesar de ser um erro de medida, ε esteja correlacionado com x. Exemplo disso é se x também for medido com erro e os erros de medida de ambas as variáveis forem correlacionados (para ilustrar, pense em uma pesquisa domiciliar em que alguns entrevistadores na dúvida arredondam variáveis para cima e outros, para baixo).

4.2.1 Erro de medida na variável explicativa

Já que falamos em erro de medida, aproveito para comentar rapidamente sobre problemas decorrentes da presença de erros de medida na variável explicativa. Em verdade, este problema pode ou não dar uma interpretação de puro erro de medida ao erro da regressão, mas além disso provoca um segundo problema em nosso exercício: viesa os coeficientes de uma regressão linear na direção de zero.

Para ver isso, vamos considerar um modelo linear do tipo:

$$\begin{aligned} y &= b_0 + b_1 x + u \\ x^* &= x + v \\ E(u|X, v) &= 0 \\ E(v|X) &= 0 \end{aligned}$$

neste caso, u é o componente não-observável original do modelo, e v é um erro de medida em x. O caso em questão é tal que nossos dados contêm apenas y e x^*, mas não x. Podemos

reescrever o modelo como:

$$y = b_0 + b_1 x^* + \underbrace{(u - b_1 v)}_{\varepsilon}$$

Se estimarmos b_1 por MQO, teremos:

$$\begin{aligned} \widehat{b}_1 &= \frac{cov\,(x^*, y)}{var\,(x^*)} = \frac{cov\,(x + v, b_0 + b_1 x + u)}{var\,(x) + var\,(v)} \\ E\left(\widehat{b}_1\right) &= \underbrace{\frac{var\,(x)}{var\,(x) + var\,(v)}}_{\in(0,1)} b_1 \\ &\Longrightarrow \left|E\left(\widehat{b}_1\right)\right| < b_1 \end{aligned}$$

4.3 ε = erro de previsão

A terceira interpretação comum para o componente aleatório do modelo é a de um erro de previsão. Consideramos neste caso um modelo do tipo:

$$y_{t+1} = \mu\,(X_t) + \varepsilon_{t+1}$$

Em geral, esta interpretação surge em casos em que a variável dependente, y, é o indicador fundamental para a tomada de decisão dos indivíduos, mas a estrutura do problema é tal que os agentes tomam uma decisão em um instante do tempo, t, com base no valor que eles esperam que y tenha no período seguinte, $t+1$. Quando esta interpretação é a mais provável, é plausível supor que $E\,(\varepsilon_{t+1}|X_t) = 0$.

Um caso clássico é o de composição de uma carteira de ativos (portfólio), onde a quantidade de cada ativo que os indivíduos gostariam de manter depende entre outras coisas do

retorno esperado daquele ativo, condicional suas características de risco. A decisão é tomada em um ponto do tempo, mas os retornos se realizam em outro ponto. Se fizermos então uma regressão de retornos de ativos observados em $t+1$ em características de risco do ativo em t (para obter o retorno esperado *condicional* em suas características de risco, que a teoria prediz que deveria ser constante entre ativos), podemos atribuir a diferença entre os retornos realizados e os retornos esperados a um erro de previsão:

$$\begin{aligned} \varepsilon_{t+1} &= R_{t+1} - E\left(R_{t+1}|X_t\right) \\ E\left(R_{t+1}|X_t\right) &= X_t'\beta \end{aligned}$$

Nesta interpretação, é natural e convincente supor que $E\left(\varepsilon_{t+1}|X_t\right) = 0$. Por quê? Porque $E\left(R_{t+1}|X_t\right)$ é a melhor previsão de R_{t+1} que os agentes poderiam fazer com as informações disponíveis X_t. Se $E\left(\varepsilon_{t+1}|X_t\right) = C\left(X_t\right) \neq 0$, então os agentes poderiam melhorar sua previsão do retorno esperado adicionando à previsão anterior $C\left(X_t\right)$.

4.4 ε = variável omitida

A última interpretação a ser listada é justamente a que será mais explorada ao longo deste curso. Segundo esta interpretação, ε deveria representar um conjunto de determinantes de y que simplesmente não está presente em nossa base de dados. Tal como x, ε apreenta alguma dispersão na população, o que justificaria o fato de que mesmo dois indivíduos observacionalmente idêncticos (isto é, com mesmas características x) possam ter y distinto.

As vantagens desta interpretação são as de que (i) caso uma nova base de dados surja,

permitindo incluir no exercício empírico variáveis anteriormente não-observadas, podemos verificar se os coeficientes das variáveis originais muda na direção que nossa teoria prediria, permitindo com que mesmo as estimações anteriores (erradas), nos ajudem a entender o fenômeno; (ii) podemos eventualmente modelar explicitamente a relação entre ε e x, caso a hipótese de exogeneidade estrita ($E(\varepsilon|X) = 0$) ou outra usada para identificar e estimar o modelo não seja válida. No exemplo usado em sala, em que y é o logarítimo do salário horário e x é a escolaridade, podemos imaginar que o coeficiente de x esteja viesado porque ε seria, por exemplo, inteligência (que ao mesmo tempo estaria "causando" salários e escolaridade, e portanto viesando para cima o coeficiente de x). Neste caso, uma solução possível é incluir no modelo uma segunda equação a ser estimada, associando explicitamente escolaridade (agora como variável dependente) com inteligência (não-observável). Além disso, se a interpretação mais provável para ε é de que seja inteligência, podemos a partir disto determinar que variável presente na base de dados poderia ser convincentemente usada como um bom instrumento para ε (antecipando um pouco o que será a estimação por variáveis instrumentais, que estudaremos no curso).

De um modo geral, ε pode tanto estar captando uma única variável omitida quanto uma função de um conjunto de variáveis omitidas. Como não observamos ε, não é possível testar a verdadeira interpretação para esta variável, mas podemos conjecturar qual seria seu verdadeiro papel e a partir disto construir nossa estratégia empírica apropriada. Para ser preciso, se nosso modelo é

$$y = \mu^* (x_1, ..., x_K, e_1, ..., e_G)$$

e usarmos a hipótese de separabilidade aditiva:

$$y = \mu (x_1, ..., x_K) + f (e_1, ..., e_G)$$

então teremos $\varepsilon = f (e_1, ..., e_G)$ (ou num contexto linear, $\varepsilon = \gamma' e$).

Note que esta interpretação não exclui a priori a possibilidade de que hipóteses tais como $E (\varepsilon | X) = 0$ sejam verdadeiras. O problema é justificar porque observamos correlações não nulas entre os diferentes componentes observáveis de x $(x_1, ..., x_K)$ e justamente ente x e ε essa relação seria zero. A menos que justifiquemos esse tratamento especial para este determinante de y (que em princípio deveria receber tratamento semelhante a x, exceto pelo fato de que observamos uns e não outros), será difícil convencer uma platéia especializada de que tal hipótese é pertinente.

5 Fontes de variação

Como repetimos desde o início, nosso foco principal no curso será buscar formas de estimar o efeito sobre y de uma variação em um particular determinante x_k, mantendo-se todos os demais determinanets x_{-k} fixos. Este é o que chamamos de efeito direto de x_k sobre y, $\partial y / \partial x_k$, ou efeito ceteris-paribus, ou em mutos casos, efeito causal. Eventualmente, podemos também estar interessados no efeito total que uma variação de x_k exerce sobre y, $dy/dx_k = \partial y / \partial x_k + \sum_{l \neq k} (\partial y / \partial x_l) (\partial x_l / \partial x_k)$. O objetivo desta subseção é discutir que tipos de variação

em y e x estão presentes nos dados e que podem ser exploradas para identificar o efeito desejado.

5.1 Variação natural

A maioria das bases de dados disponíveis apresenta variáveis dependentes, y, e explicativas, x, que possuem, na amostra, uma distribuição $f(y,x)$. Quando exploramos a variação natural para identificar $\partial y/\partial x$ (em geral aproximado por $[E(y|X=x+\Delta)-E(y|X=x)]/\Delta$), essencialmente estamos supondo que duas populações caracterizadas por diferentes valores de X (digamos, $X=x+\Delta$ e $X=x$) são em alguma medida semelhantes nas demais variáveis não-observáveis que determinam o comportamento de y. Para ilustrar, se Y for salário, X for escolaridade e ε for inteligência, num exercício em que estamos interessados em estimar o efeito da escolaridade sobre salários, então a variação natural só será válida se em alguma medida os níveis de inteligência forem semelhantes entre os diferentes grupos de escolaridade. Quão semelhantes? No caso extremo, a distribuição inteira de ε precisaria ser a mesma para todo nível x de X (hipótese de independência. Por exemplo, $\varepsilon|X \sim N(0,\sigma^2)$). Num caso intermediário, a média de ε precisaria ser a mesma para todo nível x de X ($E(\varepsilon|X)=0$, exogeneidade estrita. Uma hipótese mais fraca é de que simplesmente ε e X precisam ser não correlacionados ($E(\varepsilon X)=0$, ortogonalidade).

Caso de fato os níveis de ε sejam semelhantes entre diferentes subamostras homogêneas em X, podemos calcular a diferença entre as médias de Y nos grupos $X=x+\Delta$ e $X=x$ e dizer que esta diferença representa o efeito de X passar de x para $x+\Delta$ sobre Y. A idéia aqui é de que os grupos x e $x+\Delta$ deveriam ser bastante parecidos em tudo, exceto que um possui

valores de X maiores, e sendo esta a única diferença entre os grupos, quaisquer diferenças sistemáticas em Y nestes dois grupos deveriam ser atribuídas precisamente à variação em X.

5.2 Variação longitudinal

Quando observamos uma amostra de indivíduos ao longo do tempo, podemos associar variações de Y em relação ao que seria o padrão de cada indivíduo, a variações em X com respeito ao nível comum de X. Se por exemplo queremos medir como a quantidade de vitamina C no corpo afeta um indicador de saúde, a diferença entre a variação longitudinal e a variação natural seria a de que no primeiro caso associaríamos desvios do indicador de saúde em relação ao que seria o padrão dos indivíduos a desvios na ingestão de vitamina C em relação ao nível que os indivíduos naturalmente possuem. Já no caso da variação natural, compararíamos pessoas que naturalmente possuem mais vitamina C no corpo com outras que possuem menos desta substância, e associaríamos esta variação a diferenças entre estes dois grupos em seus indicadores de saúde.

O problema neste segundo caso é que pessoas que naturalmente possuem mais vitamina C no corpo podem também possuir diferentes níveis de características não-observáveis ε que afetam saúde do que aquelas que possuem pouca vitamina C. Neste caso, parte do efeito que em uma regressão estaria sendo atribuído à vitamina C na verdade seria causado por ε e erroneamente atribuído à vitamina C pelo fato de ε não constar na regressão e estar correlacionado com vitamina C. Se os níveis naturais de ε e vitamina C forem correlacionados ($E(\varepsilon|X) \neq 0$) mas as variações ao longo do tempo de ε forem não correlacionadas às variações de X ao longo do tempo, então poderemos medir o efeito causal de X sobre Y verificando se

em períodos em que X está acima de seu nível natural coincidem com períodos em que Y aumenta/ diminui.

Da mesma forma, a variação natural pode ser preferível, se os níveis naturais de ε e X entre indivíduos forem não-correlacionados, mas suas variações ao longo do tempo o forem.

5.3 Variação experimental (controlada)

Juntamente com a variação natural, esta se situa entre as variações mais exploradas para estimar efeitos causais. O exemplo mais comum é o do medicamento e placebo. Suponha que sua variável dependente é um indicador de saúde, e seu objetivo seja o de investigar se determinada substância afeta de forma causal este indicador. Nesta estratégia, divide-se uma determinada população-alvo aleatoriamente em dois grupos, e para um deles (dito de tratamento) se administra um medicamento (isto é, algumas pessoas ingerem a substância), ao passo que para o outro grupo (grupo de controle) se administra um placebo (pilula identica, mas com conteúdo totalmente inócuo). Finalmente, observa-se se a média do indicador de saúde no grupo tratado difere significativamente da do grupo não-tratado.

A partir desta descrição, fica nítida a diferença para a variação natural, representada no exemplo pela variação nas quantidades da referida substância que os indivíduos possuem no corpo antes do experimento iniciar (se quiséssemos explorar a variação natural, ao invés de separar aleatoriamente a população em dois grupos para só então forçar exogenamente a variação de X - a substância contida no medicamento -, dividiríamos a população em grupos que naturalmente possuem quantidades distintas da substância no corpo e mediríamos a média do indicador de saúde em cada um dos grupos).

Para muitos, a variação experimental é a que produz estimativas mais convincentes do que seria o efeito causal de X sobre Y, uma vez que por construção a variação produzida em X deveria ser não relacionada aos valores das características não-observáveis que as pessoas naturalmente possuem, e que potencialmente poderiam estar correlacionados à distribuição natural de X.

Críticas a esta abordagem em geral estão associadas a dois problemas: (i) experimentos controlados são raros em situações de interesse econômico (basta imaginar que não seria factível eleger aleatoriamente uma fração da população para receber educação, por exemplo, para que pudéssemos medir desta forma o efeito de educação sobre salários). Quando ocorrem, podem ser dificilmente replicáveis ou podem ser tão específicos que a questão a que tais experimentos permitem responder se tornem economicamente desinteressantes; e (ii) indivíduos podem reagir tanto ao objeto do experimento quanto ao desenho do experimento. Se por exemplo mudássemos o critério de seleção na universidade de uma (ou mais) prova(s) para um sorteio, diminuiríamos com isso o incentivo para que bons alunos se candidatassem e aumentaríamos o incentivo para que os menos aptos participassem. Desse modo, a estimação ex-post do efeito de ter feito universidade sobre salários explorando esta variação experimental de fato mediria um efeito causal, mas válido apenas em situações onde o método de seleção para a universidade seja o sorteio. Se depois de realizado o experimento o sistema de seleção voltasse ao de avaliações, o efeito medido no experimento não necessariamente coincidiria com o impacto que a universidade teve sobre os salários dos selecionados por vestibular.

5.4 Variação causada por instrumento (experimento natural)

A última forma de variação mencionada nestas notas é em certo sentido semelhante à anterior. A principal diferença é que neste caso, os grupos de tratamento e de controle não são necessariamente divididos de forma aleatória, mas sim de acordo com alguma variável exógena à relação existente entre Y e (X, ε). Para manter a analogia com o exemplo anterior, suponha que em determinado período tenha chovido de modo mais intenso em uma ergião específica, e que este evento tenha aumentado a concentração da substância que supostamente afetaria os indicadores de saúde dos indivíduos nos alimentos. Desse modo, o grupo de pessoas que mora na região em que choveu teria sido exposto a maior ingestão da substância por uma razão exógena ao modelo (supomos que o fato de ter chovido mais em determinada área não tem efeito direto sober a saúde, e apenas afeta este indicador por seu efeito sobre a ingestão da referida substância), e desempenharia desse modo papel semelhante ao do grupo de controle em experimento controlado. Os moradores das áreas sem chuva manteriam seus níveis naturais da substância e seriam para os propósitos de nossa análise, o grupo de controle.

De um modo geral, um instrumento (nesse exemplo a chuva) é uma variável observável Z com duas características importantes: (i) é capaz de afetar diretamente os níveis da variável explicativa de interesse, X; e (ii) não afeta os níveis das variáveis não-observáveis, ε. Ao não afetar ε, o único efeito possível de Z sobre Y ocorre através do impacto de Z sobre X. Para medir finalmente o efeito de X sobre Y, comparamos apenas a variação de X induzida por Z com a respectiva variação em Y.

6 Regressão Linear

Considere o modelo:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \varepsilon$$

Onde uma variável Y é determinada por um conjunto de variáveis observáveis $X = (X_1, ..., X_K)$ e uma variável não observável (ou função de variáveis não-observáveis), ε. Vamos por enquanto supor que vale a exogeneidade estrita, $E(\varepsilon|X) = 0$. Neste caso, o efeito direto de uma determinada variável X_k sobre y é:

$$\frac{\partial y}{\partial x_k} = \beta_k$$

O objetivo desta seção é mostrar que a estimação de β_k por mínimos quadrados tem explicitamente incorporada a idéia de medir o efeito de X_k sobre Y, mantendo-se as demais variáveis X_{-k}, ε constantes. Adicionalmente, vamos mostrar que a análise de mínimos quadrados explora a variação natural existente nos dados para quantificar β_k.

O estimador do vetor de coeficientes $\beta_{(K+1)x1} = (\beta_0, \beta_1, ..., \beta_K)$ de mínimos quadrados ordinários resolve:

$$\min_{(\beta_0, \beta_1, ..., \beta_K)} \sum_{i=1}^{N} (y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K)^2$$

que tem como condições de primeira ordem:

$$\sum_{i=1}^{N} y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K = 0$$

$$\sum_{i=1}^{N} x_{1i} \left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K \right) = 0$$

$$\vdots$$

$$\sum_{i=1}^{N} x_{Ki} \left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K \right) = 0$$

6.1 Forma matricial

Note que o sistema de equações lineares (nas incógnitas $\beta_0, \beta_1, ..., \beta_K$) pode ser escrito de forma matricial como:

$$X' \left(y - X \right) \beta = 0$$

onde:

$$X_{Nx(K+1)} = \begin{bmatrix} 1 & x_{11} & x_{K1} \\ 1 & x_{12} & x_{K2} \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{KN} \end{bmatrix} = \underbrace{[1, x_1, ..., x_K]}_{Vetores}$$

$$y_{Nx1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}; \beta_{(K+1)x1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

Fácil ver que resolvendo o sistema acima obtemos:

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

6.2 Análise da regressão em dois estágios

Considere o modelo:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \varepsilon$$

Para obter o vetor de parâmetros β, devemos resolver o sistema de equações:

$$\sum_{i=1}^{N} y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K = 0$$
$$\sum_{i=1}^{N} x_{1i}\left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K\right) = 0$$
$$\vdots$$
$$\sum_{i=1}^{N} x_{Ki}\left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K\right) = 0$$

Uma forma alternativa que fornece um elemento específico deste vetor, β_k, é proceder em dois estágios, supondo que o regressor X_k é ele mesmo uma função linear dos demais regressores, X_{-k} e de um resíduo, r_k (que representaria desse modo parte da variação de X_k não explicada pelos demais regressores X_{-k}).

$$x_k = \delta_0 + \delta_1 x_1 + ... + \delta_{k-1}x_{k-1} + \delta_{k+1}x_{k+1} + ... + \delta_K x_K + r_k$$

Posto desta forma, r_k captura a variação de X_k mantendo-se os demais regressores, X_{-k} constantes ao nível x_{-k}. Com essa estrutura, podemos agora estimar $\widehat{r}_k = x_k - \sum_{l\neq k}\widehat{\delta}_l x_l$ usando $\widehat{\delta}^{MQO}$. No segundo estágio, fazemos uma regressão linear simples de y em $\widehat{r}_k$, para obter:

$$y = \beta_k \widehat{r}_k + e$$

onde $\widehat{\beta}_k = \frac{\sum_{i=1}^{N} y_i \widehat{r}_{ik}}{\sum_{i=1}^{N} \widehat{r}_{ik}^2}$.

Proof. Seja $\widehat{x}_{ik} = \sum_{l\neq k}\widehat{\delta}_l x_l$, de modo que $\widehat{r}_{ik} = x_{ik} - \widehat{x}_{ik}$. Por construção, sabemos que $\widehat{x}_{ik}$ é ortogonal a x_l, para todo $l \neq k$, o que pode ser visto através do sistema de equações que

resolve $\widehat{\delta}^{MQO}$:

$$\sum_{i=1}^{N} x_{ki} - \delta_0 - x_{1i}\delta_1 - ... - x_{Ki}\delta_K = 0 \rightarrow \sum_{i=1}^{N} r_{ki} = 0$$

$$\sum_{i=1}^{N} x_{1i}\left(x_{ki} - \delta_0 - x_{1i}\delta_1 - ... - x_{Ki}\delta_K\right) = 0 \rightarrow \sum_{i=1}^{N} x_{1i} r_{ki} = 0$$

$$\vdots$$

$$\sum_{i=1}^{N} x_{Ki}\left(x_{ki} - \delta_0 - x_{1i}\delta_1 - ... - x_{Ki}\delta_K\right) = 0 \rightarrow \sum_{i=1}^{N} x_{Ki} r_{ki} = 0$$

Se substituirmos $x_{ik} = \widehat{x}_{ik} + \widehat{r}_{ik}$ no sistema que resolve $\widehat{\beta}^{MQO}$, temos, para a condição $\sum_{i=1}^{N} x_{Ki}\left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K\right) = 0$:

$$\sum_{i=1}^{N} \left(\widehat{x}_{ik} + \widehat{r}_{ik}\right)\left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K\right) = 0$$

$$= \sum_{i=1}^{N} \widehat{x}_{ik}\left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K\right) + \sum_{i=1}^{N} \widehat{r}_{ik}\left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K\right)$$

Analisando o primeiro termo da soma acima:

$$\sum_{i=1}^{N} \widehat{x}_{ik}\left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K\right) = \sum_{i=1}^{N} \widehat{x}_{ik}\widehat{\varepsilon}_i = \sum_{i=1}^{N}\sum_{l\neq k} \widehat{\delta}_l x_l \widehat{\varepsilon}_i$$

$$= \widehat{\delta}_0 \sum_{i=1}^{N} \widehat{\varepsilon}_i + \widehat{\delta}_1 \sum_{i=1}^{N} x_{1i}\widehat{\varepsilon}_i + ... + \widehat{\delta}_{k-1} \sum_{i=1}^{N} x_{k-1i}\widehat{\varepsilon}_i + \widehat{\delta}_{k+1} \sum_{i=1}^{N} x_{k+1i}\widehat{\varepsilon}_i + ... + +\widehat{\delta}_K \sum_{i=1}^{N} x_{Ki}\widehat{\varepsilon}_i$$

mas do sistema que resolve $\widehat{\beta}^{MQO}$, fica claro que para todo x_l, $\sum_{i=1}^{N} x_{li}\widehat{\varepsilon}_i = 0$, da mesma forma que $\sum_{i=1}^{N} \widehat{\varepsilon}_i = 0$. O termo acima é portanto uma soma de zeros.

Analisando o segundo termo, temos:

$$\sum_{i=1}^{N} \widehat{r}_{ik} \left(y_i - \beta_0 - x_{1i}\beta_1 - ... - x_{Ki}\beta_K \right)$$

$$= \sum_{i=1}^{N} \widehat{r}_{ik} \left(y_i - x_{ki}\beta_k \right) - \beta_0 \sum_{i=1}^{N} \widehat{r}_{ik} - \beta_1 \sum_{i=1}^{N} x_{1i}\widehat{r}_{ik} - ... - \beta_{k-1} \sum_{i=1}^{N} x_{k-1i}\widehat{r}_{ik} - \beta_{k+1} \sum_{i=1}^{N} x_{k+1i}\widehat{r}_{ik} - ... - \beta_K \sum_{i}$$

da solução de $\widehat{\delta}^{MQO}$, sabemos que $\sum_{i=1}^{N} \widehat{r}_{ik} = 0$ e $\sum_{i=1}^{N} x_{li}\widehat{r}_{ik} = 0$ para todo $l \neq k$, o que faz com que a equação que combina os dois termos acima se reduza a:

$$\sum_{i=1}^{N} \widehat{r}_{ik} \left(y_i - x_{ki}\beta_k \right) = 0$$

levando ao estimador:

$$\widehat{\beta}_k = \frac{\sum_{i=1}^{N} \widehat{r}_{ik} y_i}{\sum_{i=1}^{N} \widehat{r}_{ik} x_{ki}}$$

Finalmente, como $x_{ki} = \widehat{x}_{ki} + \widehat{r}_{ki}$ e $\sum_{i=1}^{N} \widehat{r}_{ik}\widehat{x}_{ki} = 0$, o denominador da expressão acima fica simplesmente $\sum_{i=1}^{N} \widehat{r}_{ik}^2$. ■

Qual a intuição para o resultado acima? Se as hipóteses de que $y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \varepsilon$ e $E(\varepsilon|X) = 0$ forem válidas, o parâmetro β_k deveria medir $\frac{\partial y}{\partial x_k}$, isto é, o efeito de x_k sobre y, mantendo-se os demais regressores, x_{-k}, constantes. Como r_k captura exatamente a variação de x_k mantendo-se x_{-k} fixo, podemos usar uma regressão simples de y em r_k para obter β_k.

6.3 Análise da regressão como estimação de desvios em relação à média

Considere a regressão múltipla de y em x_1 e D, onde D é uma variável dummy (um regressor binário):

$$y = b_0 + b_1 x_1 + b_2 D + \varepsilon$$

O sistema de equações que resolve b^{MQO} nesse caso é:

$$\begin{aligned}
\sum_{i=1}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 x_{1i} - \widehat{b}_2 D_i &= \sum_{i=1}^{N} \widehat{\varepsilon}_i = 0 \\
\sum_{i=1}^{N} x_{1i}\left(y_i - \widehat{b}_0 - \widehat{b}_1 x_{1i} - \widehat{b}_2 D_i\right) &= \sum_{i=1}^{N} x_{1i}\widehat{\varepsilon}_i = 0 \\
\sum_{i=1}^{N} D_i\left(y_i - \widehat{b}_0 - \widehat{b}_1 x_{1i} - \widehat{b}_2 D_i\right) &= \sum_{i=1}^{N} D_i\widehat{\varepsilon}_i = 0
\end{aligned}$$

Preste atenção na última equação. D_i é uma variável binária, ou seja, é 0 para alguns indivíduos e 1 para outros. A soma acima pode portanto ser representada como:

$$\begin{aligned}
\sum_{i=1}^{N} D_i\left(y_i - \widehat{b}_0 - \widehat{b}_1 x_{1i} - \widehat{b}_2 D_i\right) &= \sum_{D_i=1} y_i - \widehat{b}_0 - \widehat{b}_1 x_{1i} - \widehat{b}_2 D_i = 0 \\
&\text{ou} \\
\sum_{D_i=1} \widehat{\varepsilon}_i &= 0
\end{aligned}$$

ou (se dividirmos os dois lados por $N^{D=1}$ = número de observações com $D_i = 1$)

$$\overline{\widehat{\varepsilon}}^{D=1} = 0$$

Temos assim que neste caso não apenas a média de $\widehat{\varepsilon}$ tem que ser zero na amostra como um todo, mas também tem que ser zero na subamostra de indivíduos com $D_i = 1$. mais do que isso, a média de $\widehat{\varepsilon}$ na subamostra de $D_i = 0$ também tem que ser 0, pois a média de $\widehat{\varepsilon}$ pode ser escrita como $\overline{\widehat{\varepsilon}} = P\overline{\widehat{\varepsilon}}^{D=1} + (1-P)\overline{\widehat{\varepsilon}}^{D=0}$, onde P representa a proporção da amostra

com $D_1 = 1$ (i.e. $P = N^{D=1}/N$), e se tanto $\overline{\widehat{\varepsilon}}$ quanto $\overline{\widehat{\varepsilon}}^{D=1}$ têm que ser zero, $\overline{\widehat{\varepsilon}}^{D=0} = 0$. Voltando agora ao modelo original:

$$\begin{aligned} y &= \widehat{b}_0 + \widehat{b}_1 x_1 + \widehat{b}_2 D + \widehat{\varepsilon} \Longrightarrow \frac{1}{N^{D=1}} \sum_{D_i=1} y_i = \frac{1}{N^{D=1}} \sum_{D_i=1} \widehat{b}_0 + \widehat{b}_1 x_1 + \widehat{b}_2 D + \widehat{\varepsilon} \\ &\Longrightarrow \overline{y}^{D=1} = \widehat{b}_0 + \widehat{b}_1 \overline{x}_1^{D=1} + \widehat{b}_2 + \overline{\widehat{\varepsilon}}^{D=1} \end{aligned}$$

e de modo análogo:

$$\overline{y}^{D=0} = \widehat{b}_0 + \widehat{b}_1 \overline{x}_1^{D=0} + \overline{\widehat{\varepsilon}}^{D=0}$$

O que acontece se, para cada indivíduo, subtrairmos x e y das respectivas médias mara o grupo de D ao qual pertence (isto é, construirmos, para variáveis Z, $z - \overline{z}^{D=1}$ para quem tem $D = 1$ e $z - \overline{z}^{D=0}$ para quem tem $D = 0$)? Se imaginarmos que D é sexo, por exemplo, este procedimento seria simplesmente subtrair de z individual a média de z entre os homens, para quem for homem, e a média de z entre as mulheres, para quem for mulher.

Vejamos: para quem tem $D = 1$, sabemos que $y = \widehat{b}_0 + \widehat{b}_1 x_1 + \widehat{b}_2 + \widehat{\varepsilon}$, e que $\overline{y}^{D=1} = \widehat{b}_0 + \widehat{b}_1 \overline{x}_1^{D=1} + \widehat{b}_2 + \overline{\widehat{\varepsilon}}^{D=1}$, o quem implica:

$$\begin{aligned} y - \overline{y}^{D=1} &= \widehat{b}_1 \left(x_1 - \overline{x}_1^{D=1} \right) + \left(\widehat{\varepsilon} - \overline{\widehat{\varepsilon}}^{D=1} \right) \\ &= \widehat{b}_1 \left(x_1 - \overline{x}_1^{D=1} \right) + \widehat{\varepsilon} \end{aligned}$$

e para quem possui $D = 0$, $y = \widehat{b}_0 + \widehat{b}_1 x_1 + \widehat{\varepsilon}$, e que $\overline{y}^{D=1} = \widehat{b}_0 + \widehat{b}_1 \overline{x}_1^{D=1} + \overline{\widehat{\varepsilon}}^{D=1}$, o quem implica:

$$\begin{aligned} y - \overline{y}^{D=0} &= \widehat{b}_1 \left(x_1 - \overline{x}_1^{D=0} \right) + \left(\widehat{\varepsilon} - \overline{\widehat{\varepsilon}}^{D=0} \right) \\ &= \widehat{b}_1 \left(x_1 - \overline{x}_1^{D=0} \right) + \widehat{\varepsilon} \end{aligned}$$

Para ambos os casos temos, portanto:

$$y - \overline{y}^{D} = \widehat{b}_1 \left(x_1 - \overline{x}_1^{D} \right) + \widehat{\varepsilon}$$

e se $E(\varepsilon|X) = 0$, então $E\left(\varepsilon|x_1 - \overline{x}_1^{D}\right) = 0$ (vimos anteriormente que $E(\varepsilon|X) = 0$ implica $E(\varepsilon|g(X)) = 0$, para qualquer função de X). Logo, podemos encontrar $\widehat{b}_1$ fazendo uma regressão simples de $\widetilde{y}$ em $\widetilde{x}$, onde para uma variável aleatória z: $\widetilde{z} = z - \overline{z}^{D}$.

Temos agora uma segunda interpretação para os coeficientes de uma regressão linear múltipla: b_k mede como desvios de x_k em relação ao que seria o padrão esperado dadas suas demais características D; causam desvios em y com relação a que seria esperado dadas suas características D. Fica evidente que o tipo de variação explorada nos estimadores de mínimos quadrados para identificar b_k é, portanto, a variação natural existente nas variáveis y, x e D.

6.4 Variâncias dos estimadores

Considere primeiro um estimador de MQO de uma regressão linear simples ($y = \alpha + \beta x + \varepsilon$):

$$\begin{aligned}\widehat{\beta} &= \frac{\sum_{i=1}^{N}(x_i-\overline{x})\,y_i}{\sum_{i=1}^{N}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{N}(x_i-\overline{x})\,(\alpha+\beta x_i+\varepsilon_i)}{\sum_{i=1}^{N}(x_i-\overline{x})^2} \\ &= \beta+\sum_{i=1}^{N}\left(\frac{(x_i-\overline{x})}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right)\varepsilon_i\end{aligned}$$

Desse modo, escrevemos o estimador $\widehat{\beta}$ como uma função das variáveis aleatórias do modelo, $\{\varepsilon_i\}_{i=1}^{N}$. Como a amostra é aleatória ($\text{cov}(\varepsilon_i,\varepsilon_j)=0$, para todo $i\neq j$), a variância de $\widehat{\beta}$ é simplesmente:

$$var\left(\widehat{\beta}\right)=\sum_{i=1}^{N}\left(\frac{(x_i-\overline{x})}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right)^2 var\left(\varepsilon_i\right)$$

que no caso homocedástico fica:

$$\begin{aligned}var\left(\widehat{\beta}\right) &= \sum_{i=1}^{N}\left(\frac{(x_i-\overline{x})}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right)^2\sigma^2 \\ &= \sigma^2\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{\left(\sum_{i=1}^{N}(x_i-\overline{x})^2\right)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\end{aligned}$$

No caso de uma regressão múltipla, vimos que $\widehat{\beta}$ pode sempre ser obtido por regressão simples num segundo estágio, quando regredimos y no resíduo de uma regressão de x_k em X_{-k}:

$$var\left(\widehat{\beta_k}\right)=\frac{\sigma^2}{\sum_{i=1}^{N}\left(\widehat{r}_{ki}-\overline{\widehat{r}_k}\right)^2}=\frac{\sigma^2}{\sum_{i=1}^{N}\widehat{r}_{ki}^2}$$

Se voltarmos na definição original de $\widehat{r}_k$, temos que $\widehat{r}_{ki} = x_{ki} - \widehat{x}_{ki}$, ou seja, $\sum_{i=1}^{N} \widehat{r}_{ki}^2$ é literalmente a soma dos quadrados dos resíduos da regressão que tem x_k como variável dependente (em outras palavras: $\sum_{i=1}^{N} \widehat{r}_{ki}^2 = SQR_k$). Sabemos que, como em qualquer projeção, $SQR = SQT - SQE$, onde $SQT = \sum_{i=1}^{N} (y_i - \overline{y})^2$ e $SQE = \sum_{i=1}^{N} \left(\widehat{y}_i - \overline{\widehat{y}}\right)^2 = SQE = \sum_{i=1}^{N} (\widehat{y}_i - \overline{y})^2$. Além disso, pela definição de R^2 temos que $R^2 = SQE/SQT$, o que portanto implica que $SQR = (1 - R^2)\, SQT$. Com isso, conseguimos expressar a variância dos estimadores de uma regressão múltipla como sendo:

$$var\left(\widehat{\beta}_k\right) = \frac{\sigma^2}{(1 - R_k^2)\, SQT_k}$$

O que aprendemos com isso? Primeiramente, que quanto maior for a variância do componente não-observável, maior a variância de $\widehat{\beta}_k$. Segundo, quanto maior a variância de X_k menor a variância de $\widehat{\beta}_k$. Ambos os fatos seguem uma mesma lógica: $\widehat{\beta}_k$ captura quão importante x_k é para explicar a variação observada de y. Se grande parte da variação de y resulta da variação no componente não observado (σ^2), então $\widehat{\beta}_k$ tende a ser não significativo. Se por outro lado é a variação em x_k que explica a parte mais importante da variação de y, este fato é capturado por uma maior precisão nas estimativas de β_k. Finalmente, se x_k é muito correlacionado com os demais regressores, então grande parte do que seria seu poder explicativo na variação de y fica compartilhado com a variação dos demais regressores, e é por isso que se R_k^2 é grande, $var\left(\widehat{\beta}_k\right)$ tende a ser grande também.

7 Propriedades de distribuições normais

Notação:

1. z é uma variável aleatória distribuída segundo uma densidade normal padrão se:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2} z^2$$

2. x é uma variável aleatória distribuída segundo uma densidade normal com média μ e variância σ^2 se:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2$$

3. $\mathbf{z} = (z_1, ..., z_K)$ é um vetor de variáveis aleatórias distribuídas segundo uma distribuição normal multivariada padrão se

$$f(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^K}} \exp -\frac{1}{2} \mathbf{z}'\mathbf{z}$$

4. $\mathbf{x} = (x_1, ..., x_K)$ é um vetor de variáveis aleatórias distribuídas segundo uma distribuição normal multivariada com média $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$ e matriz de covariância $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1K} \\ \sigma_{12} & \sigma_2^2 & \\ & & \\ \sigma_{1K} & & \sigma_K^2 \end{bmatrix}$, onde σ_{kl} denota a covariância entre x_k e x_l, e σ_k^2 denota a variância de x_k, se:

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi |\boldsymbol{\Sigma}|)^K}} \exp -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

onde $|\boldsymbol{\Sigma}|$ designa o determinante da matriz $\boldsymbol{\Sigma}$.

5. Se $w = \mathbf{z}'\mathbf{z} = \sum_{k=1}^{K} z_k^2$ (se puder ser escrito como a soma dos quadrados de $z_1, ..., z_K$ variáveis normal-padrão independentes), então w se distribui segundo uma $\chi^2_{(K)}$ (qui-quadrado com K graus de liberdade).

6. Se $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $\boldsymbol{\Sigma}$ sendo uma matriz positiva-definida, então (i) existe uma matriz $\boldsymbol{\Sigma}^{1/2}$ tal que $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$, e (ii) $\left(\boldsymbol{\Sigma}^{1/2}\right)^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{z}$ (onde $\mathbf{z}$ é um vetor de variáveis aleatórias conjuntamente distribuídas segundo uma normal-padrão). Corolário: se $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então $w = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ se distribui segundo uma $\chi^2_{(K)}$ (onde K é a dimensão de $\mathbf{x}$).

7. Se $z \sim N(0,1)$; $w \sim \chi^2_{(K)}$ e $z \underline{\parallel} w$, então $t = z/\sqrt{w/K} \sim t_{(K)}$ onde $t_{(K)}$ representa uma distribuição t de Student com K graus de liberdade.

7.0.1 Linearidade

$$\text{Se } \mathbf{x}_{K\times 1} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ então } \mathbf{y}_{G\times 1} = \mathbf{A}_{G\times 1} + \mathbf{B}_{G\times K}\mathbf{x}_{K\times 1} \sim N(\mathbf{A} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$$

Corolário 1: (caso escalar) se $\mathbf{x}_{1\times 1}$, isto é, $x \sim N(\mu, \sigma^2)$, então

$$a + bx \sim N\left(a + b\mu, \mu^2\sigma^2\right)$$

Corolário 2: se duas variáveis x_1, x_2 são conjuntamente normalmente distribuídas, isto é, se $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right)$, então:

$$\begin{aligned} y &= a + bx_1 + cx_2 \sim N\left(\mu_y, \sigma_y^2\right) \\ \mu_y &= a + b\mu_1 + c\mu_2 \\ \sigma_y^2 &= b^2\sigma_1^2 + c^2\sigma_2^2 + 2bc\sigma_{12} \end{aligned}$$

(o argumento se generaliza para mais de 2 variáveis).

Corolário 3: No exemplo acima, se y for o resultado da projeção de x_2 em x_1, isto é,

$$\begin{aligned} \widehat{x}_2 &= E(x_2) + \frac{cov(x_1, x_2)}{var(x_1)}(x_1 - E(x_1)) \\ &= \mu_2 + \frac{\sigma_{12}}{\sigma_1^2}(x_1 - \mu_1) = \left(\mu_2 - \frac{\sigma_{12}}{\sigma_1^2}\mu_1\right) + \frac{\sigma_{12}}{\sigma_1^2}x_1 \end{aligned}$$

e se $r = x_2 - \widehat{x}_2$, então:

$$\begin{aligned} \begin{pmatrix} \widehat{x}_2 \\ r \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_{\widehat{x}_2} \\ \mu_r \end{pmatrix} \begin{bmatrix} \sigma_{\widehat{x}_2}^2 & 0 \\ 0 & \sigma_r^2 \end{bmatrix}\right) \\ \mu_{\widehat{x}_2} &= \mu_2 \\ \sigma_{\widehat{x}_2}^2 &= \left(\frac{\sigma_{12}}{\sigma_1}\right)^2 = \rho^2\sigma_2^2 \\ \mu_r &= 0 \\ \sigma_r^2 &= \sigma_2^2 - \left(\frac{\sigma_{12}}{\sigma_1}\right)^2 = \left(1 - \rho^2\right)\sigma_2^2 \end{aligned}$$

onde $\rho = \sigma_{12}/\sigma_1\sigma_2$ é o coeficiente de correlação entre x_1 e x_2.

7.0.2 Independência = ortogonalidade

Se $\mathbf{x}_{K\times 1} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então, para quaisquer dois elementos k, l de $\mathbf{x}_{K\times 1}$ temos que:

$$x_k \underline{\parallel} x_l \iff \sigma_{kl} = 0$$

O resultado geral é que se duas variáveis x_k, x_l são independentes, então $\sigma_{kl} = 0$ (ou seja, $x_k \underline{\parallel} x_l \implies \sigma_{kl} = 0$). No caso particular de distribuições normais, "vale a volta", isto é, $x_k \underline{\parallel} x_l \impliedby \sigma_{kl} = 0$ (ou seja, se $\sigma_{kl} = 0$ então x_k e x_l são independentes).

Se observarmos novamente o último corolário da subseção anterior, veremos que a técnica de projeção, se aplicada a variáveis que se distribuem segundo uma normal multivariada, decompõe um elemento de **x** em um componente linearmente dependente dos demais, e um componente independente dos demais.

8 Teorema do limite central, Lei dos grandes números e testes de hipótese

8.1 Teoria assintótica

Lei dos Grandes Números: Seja $x_1, x_2, ..., x_N$ uma sequência de variáveis aleatórias independente e identicamente distribuídas, com média μ e variância $\sigma^2 < \infty$. Defina a média tomada entre estas N variáveis de $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} x_n$. Então:

$$P \lim_{N\to\infty} \left(\overline{X}_N\right) = \mu$$

Teorema do Limite Central (versão escalar): Seja $x_1, x_2, ..., x_N$ uma sequência de variáveis aleatórias independente e identicamente distribuídas, com média μ e variância

$\sigma^2 < \infty$. Defina a média tomada entre estas N variáveis de $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} x_n$. Então:

$$\sqrt{N}\left(\overline{X}_N - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right)$$

ou

$$\overline{X}_N \xrightarrow{A} N\left(\mu, \frac{\sigma^2}{N}\right)$$

Teorema do Limite Central (versão multivariada): Seja $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ uma sequência de vetores aleatórios independente e identicamente distribuídos, com média $\boldsymbol{\mu}$ e matriz de covariância $\Sigma < \infty$. Defina a média tomada entre estes N variáveis de $\overline{\mathbf{X}}_N = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$. Então:

$$\sqrt{N}\left(\overline{\mathbf{X}}_N - \boldsymbol{\mu}\right) \xrightarrow{d} N\left(\mathbf{0}, \Sigma\right)$$

Teorema de Slutsky: Seja $g(.)$ uma função contínua que não tenha N entre seus argumentos. Seja $x_1, x_2, ..., x_N$ uma sequência de variáveis aleatórias independente e identicamente distribuídas, e $x^N = f(x_1, x_2, ..., x_N)$ uma função dos elementos desta sequência (por exemplo, a média) tal que $P\lim\left(x^N\right) = c$, constante. Então:

$$P\lim g\left(x^N\right) = g\left(P\lim\left(x^N\right)\right) = g(c)$$

Corolários:

Sejam $x_1, x_2, ..., x_N$ e $y_1, y_2, ..., y_N$ sequências de variáveis aleatórias independente e identicamente distribuídas, com médias μ_x, μ_y e variâncias $\sigma_x^2, \sigma_y^2 << \infty$. Sejam a, b, c constantes.

Então:

$$\begin{aligned} P\lim_{N\to\infty}\left(a\overline{X}_N\right)\left(b\overline{Y}_N\right) &= ab\mu_x\mu_y \\ P\lim_{N\to\infty} a\overline{X}_N + b\overline{Y}_N + c &= a\mu_x + b\mu_y + c \\ P\lim_{N\to\infty}\left(\frac{\overline{X}_N}{a}\right)\left(\frac{\overline{Y}_N}{b}\right) &= \frac{\mu_x\mu_y}{ab}; ab \neq 0 \end{aligned}$$

E além disso, se $x^N = f(x_1, x_2, ..., x_N)$ é tal que $x^N \xrightarrow{d} f(x)$, então $g\left(x^N\right) \xrightarrow{d} f(g(x))$.

A principal consequência para nós destes teoremas é que frequentemente gostamos de supor que nossa amostra é aleatória, o que implica que os elementos não-observáveis ε dos indivíduos que compõem nossa amostra são i.i.d. (independente e identicamente distribuídos). Como vimos também que em uma regressão com intercepto a hipótese de que $E(\varepsilon) = 0$ não limita a análise de $\partial y/\partial x$ (nosso objeto de interesse), segue que:

$$\sqrt{N}\overline{\varepsilon}_N \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

Além do mais, se supusermos que $E(\varepsilon|X) = 0$, o que implica que $E(\mathbf{x}'\varepsilon) = 0$:

$$\sqrt{N}\overline{\mathbf{x}'\varepsilon}_N \xrightarrow{d} N(\mathbf{0}, x'\Sigma x)$$

8.2 Testes de hipótese

Suponha que o modelo verdadeiro é $y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \varepsilon$. Considere inicialmente a expressão do estimador de mínimos quadrados de β :

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_0 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} \left(\frac{(x_i - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right) y_i \\ \overline{y} - \widehat{\beta}_1 \overline{x} \end{pmatrix} \rightarrow \text{Regressão linear simples}$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \rightarrow \text{Regressão linear múltipla (representação matricial)}$$

Em ambos os casos, podemos reescrever a expressão do estimador como sendo uma função linear de vetor de não-observáveis, ε :

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_0 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^{N} \left(\frac{(x_i - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right) \varepsilon_i \\ \sum_{i=1}^{N} \frac{1}{N} - \left(\frac{\overline{x}(x_i - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right) \varepsilon_i \end{pmatrix} \rightarrow \text{Regressão linear simples}$$

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon} \rightarrow \text{Regressão linear múltipla (representação matricial)}$$

Podemos agora derivar a distribuição dos estimadores em dois casos: (i) pequenas amostras, supondo que conhecemos a distribuição de ε, e (ii) grandes amostras, aproveitando o fato de que

$$\mathbf{X}'\boldsymbol{\varepsilon} = \begin{pmatrix} \sum_{i=1}^{N} \varepsilon_i \\ \sum_{i=1}^{N} x_{1i}\varepsilon_i \\ \vdots \\ \sum_{i=1}^{N} x_{Ki}\varepsilon_i \end{pmatrix} = N \begin{pmatrix} \overline{\varepsilon} \\ \overline{(x_1\varepsilon)} \\ \vdots \\ \overline{(x_K\varepsilon)} \end{pmatrix}$$

e de que podemos aplicar o Teorema do Limite Central em $\overline{\varepsilon}, \overline{(x_1\varepsilon)}, ..., \overline{(x_K\varepsilon)}$ e obter a distribuição-limite destas médias.

Seja $\boldsymbol{\Sigma}$ a matriz de covariância de $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_N)$, ou seja (obs: na notação matricial, a matriz de covariância de um vetor $\mathbf{v}$ é $cov(\mathbf{v}) = E(\mathbf{v}\mathbf{v}') - E(\mathbf{v})(E(\mathbf{v}))'$):

$$E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right)=\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N}\\ \sigma_{12} & \sigma_2^2 & & \vdots\\ \vdots & & \ddots & \vdots\\ \sigma_{1N} & \cdots & \cdots & \sigma_N^2\end{bmatrix}$$

Como vimos em sala, se a amostra for aleatória, isto é, se $\varepsilon_i \underline{\parallel} \varepsilon_j$ para todo $i \neq j$, então $\sigma_{ij}=0$ e a matriz é diagonal. Se além disso ε for homocedástico, então $\boldsymbol{\Sigma}=\sigma^2\mathbf{I}$.

Olhando agora para a distribuição de $\mathbf{X}'\boldsymbol{\varepsilon}$, sabemos que sob a hipótese de exogeneidade estrita, $E\left(\varepsilon|X\right)=0$, vale $E\left(\mathbf{X}'\boldsymbol{\varepsilon}\right)=0$. Por outro lado,

$$cov\left(\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}\right)=E\left(\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}\right)=\mathbf{X}'E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right)\mathbf{X}=\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}$$

No caso (i), se for válida a hipótese de que $\varepsilon|X\sim N\left(0,\sigma^2\right)$, então teremos:

$$\widehat{\boldsymbol{\beta}}\sim N\left(\boldsymbol{\beta},\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

ou no caso heterocedástico[3]:

$$\widehat{\boldsymbol{\beta}}\sim N\left(\boldsymbol{\beta},\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

ou ainda em uma regressão simples:

[3] Basta aplicar a propriedade da distribuição normal. Se $\boldsymbol{\varepsilon}\sim N\left(\mathbf{0},\boldsymbol{\Sigma}\right)$ e $\widehat{\boldsymbol{\beta}}$ é uma combinação linear de $\boldsymbol{\varepsilon}$ $(\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon})$.

$$\begin{aligned}
\widehat{\beta}_1 &\sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right) \rightarrow \text{homocedástico} \\
\widehat{\beta}_0 &\sim N\left(\beta_0, \sigma^2\left[\frac{1}{N}+\frac{\overline{x}^2}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right]\right)
\end{aligned}$$

No caso (ii), temos que $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)^{-1}\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{N}$**.** Pela Lei dos Grandes Números, $\left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)^{-1}$ converge em probabilidade para $E\left[(\mathbf{x}'\mathbf{x})^{-1}\right]$, ao passo que o Teorema do Limite Central nos garante que

$$\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{N} = \overline{\mathbf{X}'\boldsymbol{\varepsilon}} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i \\ \frac{1}{N}\sum_{i=1}^{N}x_{1i}\varepsilon_i \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N}x_{Ki}\varepsilon_i \end{pmatrix} = \begin{pmatrix} \overline{\varepsilon} \\ \overline{(x_1\varepsilon)} \\ \vdots \\ \overline{(x_K\varepsilon)} \end{pmatrix}$$

é tal que:

$$\begin{aligned}
\sqrt{N}\overline{\mathbf{X}'\boldsymbol{\varepsilon}} &\sim N\left[\mathbf{0}, E\left(\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}\right)\right] \\
&= N\left[\mathbf{0}, \mathbf{X}'E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right)\mathbf{X}\right] \\
&= N\left[\mathbf{0}, \mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\right]
\end{aligned}$$

e agora temos que $\widehat{\boldsymbol{\beta}}$ é uma combinação linear de $\overline{\mathbf{X}'\boldsymbol{\varepsilon}}$:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim N\left[\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right]$$

Agora que sabemos que de um jeito ou de outro $\widehat{\boldsymbol{\beta}}$ segue uma distribuição normal com

média $\boldsymbol{\beta}$ e variância $\mathbf{V}$, (a forma de $\mathbf{V}$ depende de estarmos no caso i ou ii). O passo seguinte é elaborar testes de hipótese a partir deste fato.

No caso de testes de uma única restrição linear sobre $\widehat{\boldsymbol{\beta}}$, podemos usar um teste normal ou um teste t. Isso porque sabemos que se

$$\widehat{\boldsymbol{\beta}} \sim N \left[\begin{array}{c} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{array} ; \left(\begin{array}{cccc} var(\beta_0) & cov(\beta_0,\beta_1) & \cdots & cov(\beta_0,\beta_K) \\ cov(\beta_0,\beta_1) & var(\beta_1) & & \vdots \\ \vdots & & \ddots & \vdots \\ cov(\beta_0,\beta_K) & \cdots & \cdots & var(\beta_K) \end{array} \right) \right]$$

então, para qualquer k, $\widehat{\beta}_k \sim N(\beta_k, var(\beta_k))$. O teste é imediato se $var(\beta_k)$ for conhecida, ou se estivermos fazendo um teste em grandes amostras, onde sabemos que $\widehat{var}(\beta_k)$ converge em probabilidade para $var(\beta_k)$. Se estivermos em pequenas amostras, podemos supor normalidade dos resíduos e usar um teste $\widehat{\beta}_k / \sqrt{\widehat{var}(\beta_k)} \sim t$. Neste caso, é conveniente supor homocedasticidade (se tal hipótese for plausível) de ε, caso em que $var(\beta_k) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}_{(k,k)}$ (obs: $(\mathbf{X}'\mathbf{X})^{-1}_{(k,k)}$ é a coordenadada (k,k) da matriz $(\mathbf{X}'\mathbf{X})^{-1}$), e $\widehat{var}(\beta_k) = \widehat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}_{(k,k)}$ (onde $\widehat{\sigma}^2 = (N-K-1)^{-1} \sum_{i=1}^{N} \widehat{\varepsilon}_i^2$). Essa categoria de testes engloba hipóteses do tipo $H_0 : \widehat{\beta}_k = b$ (onde b é uma constante, na maioria dos casos zero), e $H_0 : \widehat{\beta}_k \geq b$.

No caso de testes de um conjunto de G restrições lineares sobre o vetor $\widehat{\boldsymbol{\beta}}$, do tipo $\mathbf{R}_{G\times(K+1)} \widehat{\boldsymbol{\beta}}_{(K+1)\times 1} = \mathbf{r}_{G\times 1}$, se utilizarmos novamente as propriedades da distribuição normal, sabemos que sob a hipótese das restrições serem verdadeiras:

$$\begin{aligned} \mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r} &\sim N\left(0, \mathbf{R}var\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{R}'\right) \\ \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)' \left(\mathbf{R}var\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{R}'\right)^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right) &\sim \chi^2_{(G)} \end{aligned}$$

Em geral, usamos o segundo resultado para testar a hipótese enunciada, devido ao fato de que a estatística χ^2 se distribui de modo univariado (permite uma forma simples de fazer um teste, comparando a estatística de teste com uma tabela de valores críticos conhecida. Se fôssemos usar a normal multivariada para testar a hipótese, teríamos que ter uma tabela de valores críticos para uma distribuição em G dimensões, onde além de nos obrigar a ter uma tabela distinta para cada G, forçaria ainda que tivéssemos hiperplanos $G-$dimensionais de valores críticos, ao invés de um único número). A estatística de teste qui-quadrado acima é chamada de Estatística de Wald. Note que sob homocedasticidade a expressão acima fica:

$$\begin{aligned} & \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)' \left(\mathbf{R}var\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{R}'\right)^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right) \\ = & \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)' \left(\mathbf{R}\left[\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]\mathbf{R}'\right)^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right) \\ = & \sigma^{-2}\left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)' \left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right) \end{aligned}$$

Finalmente, devemos analisar o caso em que queremos testar $\mathbf{R}\widehat{\boldsymbol{\beta}} = \mathbf{r}$ em pequenas amostras e o valor de $var\left(\widehat{\boldsymbol{\beta}}\right)$ não é conhecido. Primeiramente, note que

$$\begin{aligned} \widehat{\varepsilon} &= y - \mathbf{x}\widehat{\boldsymbol{\beta}} \\ &= \varepsilon - \mathbf{x}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \end{aligned}$$

e sob as hipóteses de MQO, $x\left(\beta-\widehat{\beta}\right)$ é ortogonal a ε. O estimador da variância de ε sob homocedasticidade é:

$$\begin{aligned}
\widehat{\sigma}^2 &= (N-K-1)^{-1}\sum_{i=1}^{N}\widehat{\varepsilon}_i^2 \\
(N-K-1)\,\widehat{\sigma}^2 &= \sum_{i=1}^{N}\varepsilon_i^2-\sum_{i=1}^{N}\left[\mathbf{x}_i\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right]^2 \\
&= \sum_{i=1}^{N}\varepsilon_i^2+\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\mathbf{x}_i'\mathbf{x}_i\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \\
&= \sum_{i=1}^{N}\varepsilon_i^2+\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)
\end{aligned}$$

dividindo ambos os lados por σ^2:

$$\begin{aligned}
(N-K-1)\,\frac{\widehat{\sigma}^2}{\sigma^2} &= \sum_{i=1}^{N}\left(\frac{\varepsilon_i}{\sigma}\right)^2-\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\left(\sum_{i=1}^{N}\frac{\mathbf{x}_i'\mathbf{x}_i}{\sigma}\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \\
&= \sum_{i=1}^{N}\left(\frac{\varepsilon_i}{\sigma}\right)^2-\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\left(var\left(\widehat{\boldsymbol{\beta}}\right)\right)^{-1}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)
\end{aligned}$$

Como $\left(\frac{\varepsilon_i}{\sigma}\right)$ distribui-se segundo uma normal-padrão e $\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\left(var\left(\widehat{\boldsymbol{\beta}}\right)\right)^{-1}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ é a soma de $K+1$ normais-padrão ao quadrado, temos que o lado direito da equação acima é a soma de $N-K-1$ normais-padrão independentes, e portanto $(N-K-1)\,\frac{\widehat{\sigma}^2}{\sigma^2}$ distribui-se segundo uma $\chi^2_{(N-K-1)}$. Usando o fato de que a razão de duas variáveis qui-quadrado ajustadas pelos respectivos graus de liberdade, distribui-se segundo uma F, temos que:

$$\frac{\sigma^{-2}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{r}\right)'\left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{r}\right)/G}{(N-K-1)\,\frac{\widehat{\sigma}^2}{\sigma^2}/\left(N-K-1\right)}\sim F_{(G,N-K-1)}$$

simplificando a estatística de teste acima:

$$\frac{\left(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{r}\right)'\left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{r}\right)}{G\widehat{\sigma}^2}\sim F_{(G,N-K-1)}$$

fornece uma forma de testarmos o conjunto de restrições lineares $\mathbf{R}\widehat{\boldsymbol{\beta}}=\mathbf{r}$, no caso em que a amostra é pequena e não conhecemos σ^2. Note contudo que na construção desta estatística utilizamos a suposição de que os erros são homocedásticos, de modo que o teste é válido somente se esta hipótese também o for.

8.3 Efeitos homogêneos ou heterogêneos?

Ao longo do curso, vários são os objetos de nosso interesse, mas um se destaca acima dos demais. Em uma relação do tipo $y=\mu\left(\mathbf{x}\right)+\varepsilon$, examinamos com mais detalhe o efeito direto que uma variação em um determinado elemento de $\mathbf{x}$, x_k, provoca em y, mantendo-se os demais regressores fixos: $\partial y/\partial x_k|_{\mathbf{X}_{-1}=\mathbf{x}_{-k}}$.

No caso geral, é possível (e em grande parte das vezes plausível), que o efeito direto de x_k em y varie dependendo do ponto em que é avaliado, ou seja:

$$\begin{aligned}\left.\frac{\partial y}{\partial x_k}\right|_{\mathbf{X}_{-k}=\mathbf{x}_{-k}^*} &= \frac{\partial\mu\left(x_1^*,...,x_{k-1}^*,x_k,x_{k+1}^*,...,x_K^*\right)}{\partial x_k}\\ &\neq \frac{\partial\mu\left(x_1^{**},...,x_{k-1}^{**},x_k,x_{k+1}^{**},...,x_K^{**}\right)}{\partial x_k}\end{aligned}$$

(a derivada parcial com respeito a x_k avaliada no ponto $\mathbf{x}^*$ pode ser diferente do respectivo efeito avaliado no ponto $\mathbf{x}^{**}$).

Em casos particulares, o valor de $\frac{\partial y}{\partial x_k}$ independe do valor dos demais regressores em que é estimado. Para que isso aconteça, é necessário que:

$$\begin{aligned} \mu(x_1, ..., x_k, ..., x_K) &= \mu_0(\mathbf{x}_{-k}) + \beta_k x_k \\ \frac{\partial y}{\partial x_k} &= \beta_k \end{aligned}$$

Quando ocorre esta situação (comum no contexto de regressões lineares), dizemos que o efeito de x_k sobre y é homogêneo (no caso geral, é heterogêneo). Os ingredientes fundamentais para que os efeitos sejam homogêneos são (i) a separabilidade da função-resposta $\mu(x_1, ..., x_k, ..., x_K)$ em uma parte que não depende de x_k, $\mu_0(\mathbf{x}_{-k})$, e outra que depende, e (ii) a linearidade em x_k da parte da função-resposta que depende de x_k.

Como o conteúdo deste curso enfatiza especialmente modelos lineares, uma questão natural é se podemos em alguma medida estimar efeitos heterogêneos através de regressões lineares. A resposta é que quando estimamos modelos ditos lineares, nos referimos à linearidade em parâmetros β, e não à linearidade em x. Por exemplo, se estimarmos uma regressão linaer do logaritimo do salário (w) em educação (S) dos indivíduos, teremos:

$$\begin{aligned} \ln w &= b_0 + b_1 S + \varepsilon \\ \frac{\partial \ln(w)}{\partial S} &= b_1 \rightarrow \text{ homogêneo} \\ \frac{\partial w}{\partial S} &= \left(b_1 e^{b_0}\right) e^{b_1 S + \varepsilon} \rightarrow \text{ heterogêneo} \end{aligned}$$

Para poder analisar efeitos heterogêneos sem precisar lidar com modelos não-lineares,

a solução passa por (i) criar regressores que sejam funções das variáveis originais, e (ii) construir regressores que sejam interações de variáveis da base.

Se por exemplo acreditamos que a relação do logarítimo do salário com escolaridade é quadrática, podemos, a partir de uma base de dados com informações sobre salários e escolaridade, construir $x_1 =$ escolaridade ($x_1 = S$) e $x_2 =$ escolaridade^2 ($x_2 = S^2$), e rodar a regressão:

$$\ln(w) = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$$

Se agora quisermos saber o efeito de um aumento marginal na escolaridade sobre $\ln(w)$, devemos calcular: $\frac{\partial \ln(w)}{\partial S} = b_1 + 2b_2 S$. Note que o efeito dependerá do ponto de S do qual estamos partindo.

Por outro lado, se em nossos dados também tivermos informação sobre experiência dos indivíduos no mercado de trabalho, X, e acreditarmos que o efeito de um ano adicional de escolaridade sobre o logaritimo dos salários difere entre indivíduos com períodos diferentes de experiência, então podemos criar um regressor $x_3 = SX$, e fazer:

$$\ln(w) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$$

Agora, o impacto de um ano adicional de escolaridade sobre salários, $\frac{\partial \ln(w)}{\partial S} = b_1 + 2b_2 S + b_3 X$, será diferente tanto entre indivíduos com diferentes níveis educacionais quanto entre indivíduos com diferentes experiências profissionais.

Considere agora um problema em que nosso interesse seja na relação entre y e x, mas onde

y também é função de uma variável dummy, D. Por exemplo, se $y = \ln(w)$ e $x = S$, podemos imaginar que D seja o sexo do indivíduo. Desse modo, admitimos que salários difiram entre homens e mulheres, bem como entre níveis educacionais. No caso geral (impondo apenas separabilidade aditiva entre observáveis e não-observáveis), $y = \mu(S, D) + \varepsilon$. Restringindo nossa atenção a modelos lineares em parâmetros, podemos ter:

$$\begin{aligned} y &= b_0 + b_1 S + \varepsilon \\ y &= b_0 + b_1 S + b_2 D + \varepsilon \\ y &= b_0 + b_1 S + b_2 D + b_3 SD + \varepsilon \end{aligned}$$

No primeiro caso, o efeito é homogêneo em S, e a relação entre y e S no plano $y \times S$ pode ser representada por uma reta com intercepto b_0, e inclinação b_1.

No segundo caso, o efeito direto de S em y continua homogêneo, mas a relação entre y e S no plano $y \times S$ deve agora ser representada por duas retas com mesma inclinação, b_1, mas interceptos distintos (a subamostra $D = 0$ é representada por uma reta com intercepto b_0, e a subamostra $D = 1$ é representada por uma reta com intercepto $b_0 + b_3$).

No último caso, o sexo dos indivíduos afeta tanto o intercepto quanto a inclinação das retas. Neste caso os efeitos da escolaridade são heterogêneos. A partir dessa formulação, e se nosso objetivo for verificar se salários são diferentes para homens e mulheres, podemos testar tanto se o salário médio de um indivíduo sem escolaridade varia entre homens e mulheres ($H_0 : b_0 = 0$), quanto testar se salários de homens e mulheres diferem porque os retornos à escolaridade diferem entre homens e mulheres ($H_0 : b_1 = 0$). Finalmente, podemos

testar se as equações de salário são de um modo glbal diferentes entre homens e mulheres ($H_0 : b_0 = b_1 = 0$). Nos dois primeiros casos, o teste apropriado é uma estatística t ou normal, ao passo que no segundo caso, uma F ou qui-quadrado.

Suponha então que nosso inetresse seja testar se a equação de salários difere entre homens e mulheres, mas que um pesquisador o critique pela estratégia de estimar uma equação com interações do tipo:

$$\begin{aligned} y &= b_0 + b_1 S + b_2 D + b_3 SD + \varepsilon \\ H_0 &: b_2 = b_3 = 0 \end{aligned}$$

e ao invés disso proponha que você estime duas equações separadamente para homens e mulheres e teste se os coeficientes de intercepto e inclinação são iguais entre equações, isto é:

$$\begin{aligned} y &= b_0^0 + b_1^0 S + \varepsilon, \text{ se } D = 0 \text{ (homens)} \\ y &= b_0^1 + b_1^1 S + \varepsilon, \text{ se } D = 0 \text{ (mulheres)} \\ H_0 &: \left(b_0^0, b_1^0\right) = \left(b_0^1, b_1^1\right) \end{aligned}$$

Um resultado interessante é que as duas estratégias são exatamente equivalentes. ais do que isso, é possível mostrar que se estimarmos primeiramente uma equação do primeiro tipo e em seguida duas equações separadas para homens e mulheres:

$$\begin{aligned}
\widehat{b}_0 &= \widehat{b}_0^0 \\
\widehat{b}_1 &= \widehat{b}_1^0 \\
\widehat{b}_2 + \widehat{b}_0 &= b_0^1 \\
\widehat{b}_1 + \widehat{b}_3 &= b_1^1
\end{aligned}$$

Para ver este resultado, considere o sistema de equações que resolve MQO para $y = b_0 + b_1 S + b_2 D + b_3 SD + \varepsilon$:

$$\begin{aligned}
\sum_{i=1}^{N} \widehat{\varepsilon}_i &= 0 \rightarrow \sum_{i=1}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i = 0 \\
\sum_{i=1}^{N} \widehat{\varepsilon}_i S_i &= 0 \rightarrow \sum_{i=1}^{N} \left(y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i \right) S_i = 0 \\
\sum_{i=1}^{N} \widehat{\varepsilon}_i D_i &= 0 \rightarrow \sum_{i=1}^{N} \left(y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i \right) D_i = 0 \\
\sum_{i=1}^{N} \widehat{\varepsilon}_i D_i S_i &= 0 \rightarrow \sum_{i=1}^{N} \left(y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i \right) D_i S_i = 0
\end{aligned}$$

Note que $\widehat{\varepsilon}_i D_i = 0$ se $D_i = 0$, e $\widehat{\varepsilon}_i D_i S_i = 0$ se $D_i = 0$, o que implica que o sistema acima pode ser escrito como:

$$\begin{aligned}
\sum_{i=1}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i &= 0 \\
\sum_{i=1}^{N} \left(y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i \right) S_i &= 0 \\
\sum_{\substack{i=1 \\ D_i=1}}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i &= 0 \\
\sum_{\substack{i=1 \\ D_i=1}}^{N} \left(y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i \right) S_i &= 0
\end{aligned}$$

Adicionalmente, note que

$$\begin{aligned}
&\sum_{i=1}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i \\
= &\sum_{\substack{i=1 \\ D_i=1}}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i + \\
&\sum_{\substack{i=1 \\ D_i=0}}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i
\end{aligned}$$

o que implica que:

$$\sum_{\substack{i=1 \\ D_i=0}}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 S_i - b_2 D_i - b_3 S_i D_i = 0$$

e analogamente:

$$\sum_{\substack{i=1 \\ D_i=0}}^{N} \left(y_i - \widehat{b}_0 - \widehat{b}_1 S_i - b_2 D_i - b_3 S_i D_i \right) S_i = 0$$

Porém, nessas duas equações sabemos que $D_i = 0$ para todos os indivíduos. Se usarmos essa condição:

$$\begin{aligned}
\sum_{\substack{i=1\\D_i=0}}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 S_i &= 0 \\
\sum_{\substack{i=1\\D_i=0}}^{N} \left(y_i - \widehat{b}_0 - \widehat{b}_1 S_i\right) S_i &= 0
\end{aligned}$$

e se resolvermos o sistema acima para $\widehat{b}_0, \widehat{b}_1$, teremos exatamente o estimador de MQO de $\widehat{b}_0^0, \widehat{b}_1^0$ que obteríamos se fizéssemos uma regressão de y contra S apenas na subamostra $D = 0$.

Se agora usássemos as equações

$$\begin{aligned}
\sum_{\substack{i=1\\D_i=1}}^{N} y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i &= 0 \\
\sum_{\substack{i=1\\D_i=1}}^{N} \left(y_i - \widehat{b}_0 - \widehat{b}_1 S_i - \widehat{b}_2 D_i - \widehat{b}_3 S_i D_i\right) S_i &= 0
\end{aligned}$$

e a informação de que $D_i = 1$, teríamos:

$$\begin{aligned}
\sum_{\substack{i=1\\D_i=1}}^{N} y_i - \left(\widehat{b}_0 + \widehat{b}_2\right) - \left(\widehat{b}_1 + \widehat{b}_3\right) S_i &= 0 \\
\sum_{\substack{i=1\\D_i=1}}^{N} \left[y_i - \left(\widehat{b}_0 + \widehat{b}_2\right) - \left(\widehat{b}_1 + \widehat{b}_3\right) S_i\right] S_i &= 0
\end{aligned}$$

e se resolvermos o sistema acima para $\left(\widehat{b}_0 + \widehat{b}_2\right)$ e $\left(\widehat{b}_1 + \widehat{b}_3\right)$, teremos exatamente os estimadores de MQO de $\left(\widehat{b}_0^1, \widehat{b}_1^1\right)$.

O resultado geral é o de que, sempre que tivermos uma variável dummy em nossa regressão, parâmetros obtidos através da interação da dummy com TODOS os demais re-

gressores (incluindo o intercepto) e parâmetros obtidos através de regressões separadas nas subamostras $D = 0$ e $D = 1$ possuem uma relação linear 1-1 entre si. Mais do que isso, testar se todos os coeficientes das duas regressões no último caso são idênticos equivale a testar se conjuntamente todos os parâmetros envolvendo interações da dummy com os demais regressores são iguais a zero no primeiro caso.

8.3.1 Aplicação 1: Teste conjunto se todos os parâmetros são nulos

Uma aplicação direta do teste enunciado acima é o da hipótese conjunta de que todos os parâmetros são conjuntamente iguais a zero. Neste caso:

$$\begin{aligned} \mathbf{R} &= \mathbf{I}_{K+1} \text{ (matriz identidade } (K+1) \times (K+1) \text{)} \\ \mathbf{r} &= (0, ..., 0)' \end{aligned}$$

Este teste checa a hipótese de que de fato nosso modelo ajuda a explicar parte da variação de y, e adicionalmente complementa a análise a respeito de quais regressores realmente importam na análise. Em particular, vimos que no caso homocedástico uma das formas de escrever a variância de um determinado estimador $\widehat{\beta}_k$ é:

$$var\left(\widehat{\beta}_k\right) = \frac{\sigma^2}{SQT_k\left(1 - R_k^2\right)}$$

onde $SQT_k = \sum_{i=1}^{N}\left(x_{ik} - \overline{x}_k\right)^2$ é a soma dos quadrados "totais" de uma regressão que tivesse x_k como variável dependente e $\mathbf{x}_{-k} = (x_1, ..., x_{k-1}, x_{k+1}, ..., x_K)$ como variáveis explicativas; e R_k^2 fosse o resíduo dessa regressão. Ora, sabemos que se os regressores são muito correla-

cionados, isto é, se R_k^2 é alto, a variância de $\widehat{\beta}_k$ tende a ser grande e aumenta a probabilidade de não rejeitarmos H_0, mesmo quando o verdadeiro parâmetro é de fato diferente de zero. A razão é que nesse caso vários regressores estão explicando a mesma parte da variação de y, e individualmente é possível que nenhum se sobressaia. No entanto, se isso for verdade, provavelmente a hipótese de que conjuntamente os regressores expliquem parte da variação de y será confirmada num teste que rejeite a hipótese de que conjuntamente seus respectivos coeficientes sejam nulos.

8.3.2 Aplicação 2: Teste de Chow

9 Parte 2: dados longitudinais